Control and Management of Optical Infrastructures

Wessing, Henrik

Publication date:
2006

Document Version
Publisher's PDF, also known as Version of record

Henrik Wessing

Control and Management of Optical Infrastructures

Ph.D. Thesis

COM•DTU

September, 2006

Abstract

This thesis addresses selected controlling aspects of optical infrastructures. For all the topics, solutions are presented, which can be used to support the realisation of future optical networks.

Packet-to-packet power equalisation in combination with optical regenerations is a prerequisite for developing flexible optical packet switched networks. It is discussed, where, in an optical packet switch, such equalisation should take place, and it is argued that controlling the gain of a cross gain modulator before a regenerator provides the best results. In combination with digital control electronics, the equaliser equalises packet-to-packet power variations up to 9 dB with an insignificant power penalty.

A novel scheme for header processing in optical packet switched networks without header modification is proposed and thoroughly analysed. The scheme obsoletes complex header erasure and resynchronisation at the expense of a slightly larger packet header. The scalability of the scheme is analysed, and it is argued that it can be implemented in even very large networks up to 200 optical nodes without significant scalability problems.

In the thesis, the use of directly modulated lasers of 10 Gbit/s in fibre access networks is evaluated. A 1310 and a 1550 nm laser are evaluated and both are found viable for future integration into cost-efficient and highly integrated fibre to the home equipment.

Finally, the integration of telecommunications and research optical infrastructures is discussed, and the concepts for providing application driven bandwidth allocation in these integrated infrastructures are proposed. This also offers a couple of parameters as “handles” for the network operator to control the level of dynamics in his network.

Most of the covered topics are related to the two European IST projects: DAVID and MUPBED. These are detailed in the thesis.

Resumé

Denne afhandling behandler udvalgte kontrolaspekter ved optiske infrastrukturer. For alle emnerne præsenteres løsninger, der kan finde anvendelse ved realiseringen af fremtidens optiske netværk.

I forbindelse med optisk regenerering er udligning af effektniveauer mellem på hinanden efterfølgende pakker en forudsætning. Dette er nødvendigt for at kunne udvikle optiske pakkekoblede netværk. Det er diskuteret hvor i en optisk pakke switch en sådan udligning bør finde sted, og der argumenteres for, at styring af forstærkningsniveauet af en modulator (XGM) giver de bedste resultater. Med digital styringselektronik er det således muligt at udligning effektforskelle mellem to pakker på op til 9 dB med en ubetydelig forringelse af signalet til følge.

En nytænkt metode til at behandle forsendelsesoplysningerne i optiske pakkekoblede netværk uden at skulle rette i dem er foreslået og analyseret. Metoden overflødiggør kompleks sletning og resynkronisering af oplysningerne på den optiske pakke på bekostning af et forøget adressefelt. Skalerbarheden af metoden analyseres og der argumenteres for at denne kan implementeres i endog store netværk op til 200 knuder uden betydelige problemer med hensyn til skalerbarheden.

I afhandlingen undersøges brugen af 10 Gbit/s direkte modulerede lasere i tilgangsnetværk. Såvel en 1550 som en 1310 nm laser er undersøgt og begge er fundet relevante for fremtidig brug med henblik på deres potentielle prisniveau i forbindelse med fremtidig masse-integration i FTTH udstyr.

Sluttelig diskuteres integrationen af optiske infrastrukturer fra såvel teleoperatører som forskningsnet, og der foreslås fremgangsmåder for at tilbyde applikations-drevne båndbredde-allokeringer i disse integrerede infrastrukturer. Dette tildeler netværksoperatøren nogle parametre som kan benyttes som håndtag for at styre graden af dynamik i operatørens netværk.

De fleste af de behandlede emner er relateret til de to europæiske IST projekter, DAVID og MUPBED, som yderligere er beskrevet i afhandlingen.

Acknowledgements

First, I would like to thank my Professor Lars Dittmann for his supervision of my project including countless supervisor meetings. Also thanks for letting me work in the Networks group at COM-DTU, which without doubt is a unique research environment. In this relation I would like to thank all my colleagues for fruitful discussions on technical as well as recreational and political issues.

I would also like to thank all the partners in the European projects DAVID and MUPBED.

Especially, I would like to thank Bruno Lavigne and Dominique Chiaroni from Alcatel, France for the collaboration with the integration of the power equaliser and the regenerator. The time spent at Alcatel in 2003 was fruitful in several ways.

The Key Identification Scheme was developed based on one of those rare moments when an idea comes to you after the famous glass of red wine. Thanks to my former colleagues Henrik Christiansen and Tina Fjelde for helping me to mature, analyse and publish the idea.

Thanks to Ying Yan, Brian Mortensen and Michael Berger for the close collaboration with the activities in the MUPBED project and dynamicity modelling.

A special thank is expressed to Brian Sørensen, who heavily supported me with the development of electronics for the power equalisation and FTTH system.

Also thanks to my office mate Tue Lyster for playing loud music at the office and numerous rude comments, while listening to my occasional chit-chat.

Thanks again to Michael Berger and Henrik Christiansen for proofreading and commenting on this thesis.

Kgs. Lyngby, Denmark — Henrik Wessing

September, 2006

Publications

[Publ. 1] **H. Wessing**, T. Fjelde, H. Christiansen and L. Dittmann, "Novel scheme for efficient and cost-effective forwarding of packets in optical networks without header modification", *Proc. of Optical Fiber Communication Conference (OFC 2001)*, Anaheim, USA.

[Publ. 2] T. Fjelde, M. L. Nielsen, **H. Wessing**, D. Wolfson, "Wavelength conversion and optical logic" *DOPS-NYT,* vol 16(2), 2001.

[Publ. 3] T. Fjelde, D. Wolfson, A. Kloch, M. L. Nielsen, and **H. Wessing**, "SOA-based functional devices", *ITCom2001*, Denver 2001.

[Publ. 4] **H. Wessing**, H. Christiansen, T. Fjelde and L. Dittmann, "Novel Scheme for Packet Forwarding Without Header Modification in Optical Networks", *IEEE Journal of Lightwave Technology*, vol. 20(8), pp. 1277-1283, August 2002.

[Publ. 5] H. Christiansen, T. Fjelde and **H. Wessing**, "Novel Label Processing Schemes for MPLS", *Optical Networks Magazine*, vol. 3(6), pp.63-69 (2002).

[Publ. 6] H. Christiansen and **H. Wessing**, "Modeling GMPLS and Optical MPLS Network", *Procedings of 10th International Conference on Telecommunication (ICT'2003)*, Papeete, Tahiti, French Polynesia.

[Publ. 7] **H. Wessing**, "Delay reduction with provision based fair scheduling assuring QoS", *7th WSEAS International Conference on Communications*, Corfu, Greece, 2003.

[Publ. 8] B. Lavigne, **H. Wessing**, E. Balmefrezol, B. M. Sørensen and O. Leclerc, "Fast packet-to-packet power equaliser based on the association of a SOA and low cost control electronics for optical 3R regeneration: Concept demonstration and principle of validation at 10 Gbit/s", *Photonics in Switching 2003 (PS'2003)*, Versailles, France, 2003.

[Publ. 9] **H. Wessing**, B. Lavigne, B. M. Sørensen, E. Balmefrezol and O. Leclerc, "Combining control electronics with SOA to equalize packet-to-packet power variations for optical 3R regeneration in

optical networks at 10 Gbit/s", *Optical Fiber Communication Conference 2004 (OFC'2004)*, Los Angeles, USA, 2004.

[Publ. 10] D. Chiaroni, A. Dupas, E. Dutisseuil, B. Lavigne, **H. Wessing**, B. Mortensen, M. Berger, L. Dittmann, H. Linardakis, A. Salis, A. Stavdas, W. Lautenschlaeger, J. Karstaedt, L. Dembeck, G. Eilenberger, "Optical packet switching solutions for the metro and the backbone: Main issues of the DAVID project demonstration" *Proceedings of European Conference on Optical Communications (ECOC 2004),* Stockholm, Sweden.

[Publ. 11] **H. Wessing**, "MPLS transportation over optical switched networks" (Invited), *Proceedings of 4th (G)MPLS Workshop.* Girona, Spain, April 2005.

[Publ. 12] **H. Wessing**, "On Demand Network Resource Provisioning in Heterogeneous Multi Domain Networks: MUPBED Work Package 2", *Poster at MUPBED booth at Broadband Europe 2005*, Bordeaux, France, December 2005.

[Publ. 13] **H. Wessing,** "The European IST project MUPBED: Integration of Applications and Network Control Layer" *Proc. of TERENA Networking Conference (TNC'2005)*, Poznan, Poland, 2005.

[Publ. 14] **H. Wessing**, Y. Yan and M. S. Berger, "Modelling direct application to network bandwidth provisioning for high demanding research applications", *Proc. of 5th International Conference on Applied Information and Communication (AIC)*, Malta, 2005.

[Publ. 15] **H. Wessing**, "Interfacing applications with network control plane" (Invited), *Joint MUPBED and NOBEL Workshop* [online] http://www.ist-mupbed.org/Joint-NOBEL.html, Turin, Italy, November 2005.

[Publ. 16] **H. Wessing**, Y. Yan and M. S. Berger, "Simulation based analysis on dynamic resource provisioning in optical networks using GMPLS technologies", *WSEAS Transaction on Communications*, vol5(1), pp. 9-16, ISSN: 1109-2742, January 2006

[Publ. 17] **H. Wessing**, B. B. Mortensen, L. Dittmann, M. S. Berger: "Dynamic Bandwidth Allocation in MUPBED". Poster. *In Proc. of TERENA Networking Conference (TNC'2006)*, Catania, Italy, 2006.

[Publ. 18] Y. Yan, **H. Wessing,** M. Berger and L. Dittmann, "Prioritized OSPF-TE Mechanism for Multimedia Applications in MPLS Networks", *V Workshop in G/MPLS Networks*, Girona, Spain, March 2006.

[Publ. 19] Y. Yan, **H. Wessing** and M. Berger, "Bidirectional RSVP-TE for Multimedia Applications in GMPLS Networks", *In proc. of Optical Network Design and Modelling (ONDM2006)*, May 2006.

[Publ. 20] **H. Wessing**. "MUPBED point of view on application driven optical networks", *In proc. of 11th European Conference on Networks and Optical Communications (NOC 2006)*, (Invited) p. 471-478, Berlin, July 2006.

List of Abbreviations

3R	Reamplification, Reshaping and Retiming
ADC	Analog to Digital Converter
ADP	Avalanche Photo Diode
AF	Adaptation Function
APD	Avalanche Photo Diode
API	Application Programming Interface
APON	ATM PON
ASON	Automatically Switched Optical Network
AWG	Arrayed Waveguide Grating
AWGR	Arrayed Waveguide Grating Router
BER	Bit Error Rate
BMT	Burst Mode Transceiver
CDR	Clock and Data Recovery
CIDR	Classless Inter Domain Routing
CO	Central Office
CR-LDP	Constraint Based Label Distribution Protocol
CRT	Chinese Remainder Theorem
CW	Continuous Wave
CWDM	Coarse WDM
DAC	Digital to Analog Converter
DAVID	Data And Voice Integration over DWDM
DCF	Dispersion Compensated Fibre
DFB	Distributed Feedback
DRR	Deficit RR
DWDM	Dense Wavelength Division Multiplexing
EA	Electro Absorption
EDFA	Erbium Dope Fibre Amplifier

E-LSR	Edge LSR or Egress LSR
E-NNI	External NNI
EPON	Ethernet Passive Optical Networks
FE	Forwarding Engine
FEC	Forwarding Equivalence Class
FEC	Forwarding Error Correction
FIF	Forwarding Information Field
FIFO	First In First Out
FPGA	Field Programmable Gate Array
FTTH	Fibre to the Home
GbE	Gigabit Ethernet
GMPLS	Generalised Multi-Protocol Label Switching
HDTV	High Definition TV
IETF	Internet Engineering Task Force
IIR	Infinite Response
ILM	Integrated Laser Module
I-LSR	Ingress LSR
I-NNI	Internal-NNI
IP	Internet Protocol
IP	In Phase
IPDR	Input Power Dynamic Range
KIS	Key Identification Scheme
LA	Limiting Amplifier
LDP	Label Distribution Protocol
LPM	Longest Prefix Match
LSA	Link State Advertisements
LSP	Label Switched Path
LSR	Label Switched Router
MAC	Medium Access Control
MAN	Metropolitan Area Network
MEMS	Micro Electro Mechanical System
MPLS	Multi-Protocol Label Switching

MPλS	Multi-Protocol Lambda Switching
MUPBED	Multi-Partner European Test Beds for Research Networking
MZI	Mach-Zender Interferometer
NMI	Network Management Interface
NNI	Network Network Interface
NRR	Network Service Provider
NRZ	Non Return to Zero
NSR	Network Service Requester
NT	Network Terminal
OIF	Optical Internet Forum
OLT	Optical Line Terminal
OMPLS	Optical MPLS
ONT	Optical Network Terminal
ONU	Optical Network Unit
OOP	Out Of Phase
OPADM	Optical Packet Add and Drop Multiplexer
OPR	Optical Packet Router
OPS	Optical Packet Switching or Optical Packet Switch
OSNR	Optical Signal to Noise Ratio
OSPF	Open Shortest Path First
OSPF-TE	OSPF-Traffic Engineering
OXC	Optical Cross Connect
PFS	Provision-based Fair Scheduling
PON	Passive Optical Network
POTS	Plain Old Telephone System
PRBS	Pseudo Random Bit Sequence
QoS	Quality of Service
RNS	Residue Number System
RR	Round Robin
RSVP	Resource Reservation Protocol
RZ	Return to Zero
SNR	Signal to Noise Ratio

SOA	Semiconductor Optical Amplifier
SOAP	Simple Object Access Protocol
SRR	Surplus RR
TCP	Transport Control Protocol
TE	Traffic Engineering
TIA	Trans-Impedance Amplifier
T-LSR	Transit LSR
TMB	Traffic Manager Board
UNI	User Network Interface
VHDL	VHSIC Hardware Description Language
VHSIC	Very High Speed Integrated Circuit
VLSI	Very Large Scale Integration
VoD	Video on Demand
WAN	Wide Area Network
WDM	Wavelength Division Multiplexing
XGM	Cross Gain Modulator
XML	Extensible Markup Language
XPM	Cross Phase Modulation

Table of contents

1. Introduction

The stock dealers will agree that there have been serious up and down turns for companies with activities within Internet, data and telecommunications the last decade. Before Summer 2000, everything looked bright; too bright one would say knowing too well that the enormous investments in the dot.com era were not really supported by the customers. The story is well known, the bubble at the stock marked collapsed, several companies went bankrupt and the future for the telecom operators, service providers, and fibre and equipment manufacturers became clouded.

One result of the investments in the last decade is that raw fibres or dark fibres have been very widely deployed, and the number of fibres deployed often makes it more cost-efficient to use the neighbouring unused fibre compared to implement sophisticated techniques for increasing the capacity of the first. The abundant available transport capacity allows for services that users would otherwise not be willing to pay for. In this way, the available capacity triggers new services rather than upcoming demanding services initiate deployment of new capacity.

If the capacity is there, and the willingness to invest is reduced compared to the pre-bubble period, what are then the triggers for the future of telecom and related industries? Is it worthwhile to continue research in optical state of the art components and devices?

While the capacity to a large extend is there, the main focus for the research today is to use and manage the available capacity as cost-efficient as possible. This applies for all aspects of the network from the physical layer up to the top level administration and interoperability of several network domains.

In addition, high demanding services have special quality requirements to the transport network. These services can be real time interactive services like *Video on Demand* (VoD), video conferencing and other high bandwidth delay sensitive applications and services.

In other words, it is important to discard any ideological views of electrical and optical technologies and consider the network from applications to the transmission media. A holistic and pragmatic network approach is required.

This thesis addresses *control and management of optical infrastructures*, where optical infrastructures in principle cover optical fibre, optical wavelengths and optical components. The main focus in all parts of the thesis is on integration and cost-efficient solutions for future networks by proper electronic control and management.

Anyway, it is useful to be aware of the development in the optical infrastructures. The optical communication explodes in the last decades of last century with the development of low loss single mode fibres and the *Erbium Doped Fibre Amplifier* (EDFA) [1] enabling long haul communication. *Wavelength Division Multiplexing* (WDM) further increases the capacity of the fibre using several independent wavelengths. While the first implementation of optics only considers the point to point transmission, the bit rate increases and it becomes attractive to use optics for the signal processing in the network nodes [2] because of the potential lower power consumption of optics. *Micro Electro Mechanical Systems* (MEMS) [3] is introduced with (de)multiplexers to implement *Optical Cross Connects* (OXC) with reconfiguration times in the msec. range. Adding optical wavelengths converters [Publ. 2] as part of the OXC provides flexibility to the selection of wavelengths resulting in an increased overall utilisation of the network wavelengths. One of the next big steps is the optical packet switched network, where the data is transmitted as small packets and the address of the header is read electronically, while the payload is transported transparently through the network. The latter requires components and optical devices not yet mature or at least far from commercially available.

While it is doubtless that the optical technologies are superior for transmission, it is more questionable, whether they are competitive with their electronic counterparts for, e.g., optical packet switching.

Hence, this thesis addresses both a selection of electronic control topics for optical devices for future optical packet switched networks, and the electronic management of a mature network with optical channels. These two different control and management challenges are carried out in close collaboration with the two European IST projects DAVID [4] and MUPBED [5] .

Therefore, some selected topics are considered and their similarity is that they all focus in the integration of optics and electronics indicating cost-efficient solutions.

Hence, chapter 2 highlights selected control issues for the optical infrastructures in the core and access network, although the motivation for using optics in these two parts of the network is quite different. In addition, the *Multi Protocol Label Switching* (MPLS) framework is introduced as this is

used heavily as main network architecture in several following chapters. Finally, the management of optical networks through the *Automatically Switched Optical Network* (ASON) specifications for multi domains network is described.

Chapter 3 addresses the development of a power equalisation concept that is integrated with an all-optical regenerator. This enables packet to packet equalisation and regeneration in optical packet switched networks. The chapter also provides an overview of the European IST project DAVID, which the work is part of.

In chapter 4, it is demonstrated how a mathematical algorithm in combination with some control electronics significantly simplifies the requirements to the optical layer for optical packet switching. More specifically, a novel scheme is developed, which is used to avoid optical header modification, i.e., erasing and resynchronisation of an optical header to the transparently transmitted payload. The scalability of the scheme is thoroughly analysed and suggestions to the implementation in hardware is provided.

Where the focus in chapter 3 and 4 is on optical packet switching in the core network, the focus in chapter 5 is moved to the access network. Here the importance of cost is indicated and the parameters for cost-efficiency are analysed. This includes analysis of the lasers, optical components, receivers and electronics for future Fibre to the Home systems at high speed. In addition, a case study is described in which a 2x 10 Gbit/s bidirectional system is implemented and used to evaluate the performance of the electronic and optical FTTH components.

Again, in chapter 6, the focus changes from the physical layer to the management layer and administration of optical wavelengths and circuits in heterogeneous multi domain networks. The chapter describes the vertical integration from the application layer to the physical layer, i.e., how special high demanding applications can request for resources in the circuit layer. Furthermore, the level of dynamics for a dynamic optical infrastructure is evaluated through simulations.

Finally, chapter 7 concludes the thesis.

2. Controlling Optical Networks

The transition from pure optical transmission into optical networking is widely expected to bring the costs down while allowing for almost unlimited capacity. The thesis will highlight some specific challenges in the control of optical infrastructures and provide suggestions to the solutions.

This chapter provides the necessary background information to understand the context and the dependence of the selected topics, which are dealt with throughout the thesis. First, in section 2.1, the different understandings of optical infrastructures are considered. It is described how the control of wavelengths in multi-domain networks and the control of bias current to a laser all relate to optical infrastructures. Following, in section 2.2 and 2.3, the MPLS and ASON technologies are introduced as they are used as the main networking architectures throughout the thesis.

2.1. Control topics in core and access networks

Control and management of an optical infrastructure is a very broad topic, and it is thus not in the scope of this thesis to cover all details within optical communication, optical switching and optical networking. However, some selected topics are discussed in the thesis. In Fig. 2.1, the communication between a client and server passes three networks. First, the client communicates with the local access network before the communication is passed through the core network destined for the access network, which provides connectivity to the requested server.

In order to reduce the cost per bit and allow for less power consuming core node design, while increasing the network capacity, optics have been introduced in both the access and core network domain.

While the raw transmission of data has been done with optical fibres for decades, the core network nodes are still electronic, and optical to electronic conversion is usually required. The main focus within a number of projects has been the introduction of optical transparent end to end connections in the packet domain, thus combining the huge capacity of optical transmission with the flexibility of packet switching. The objective of these enor-

mous efforts has been to develop the *Optical Packet Switch* (OPS) with functionalities as described in section 2.1.1.

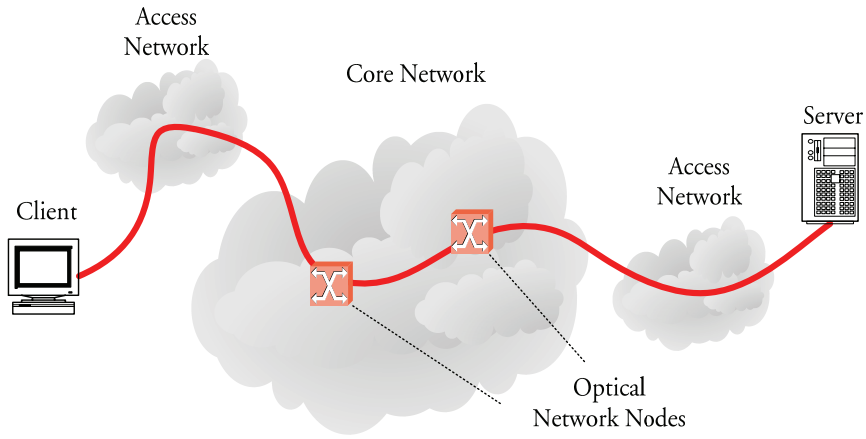


Fig. 2.1: Different types of optical networks. The red line denotes a communication between a client and a server.

In the access network the requirements for high bandwidth residential services like High Definition TV are still increasing, and it is expected that the capacity of the copper connections will not continuously be able to follow the demand. Optics is introduced at the transmission level to increase the capacity over the "last mile", which is further introduced in section 2.1.2 and in chapter 5.

In addition to the device control, the thesis also considers selected areas in the control and management of optical circuits within heterogeneous multi-domain networks. While this part is less lab-oriented its importance should not be under-estimated and the main objective is to provide dynamicity in existing optical networks, while integrating the network control with the requirements of high demanding applications. This is further introduced in section 2.1.3 and detailed in chapter 6.

2.1.1. Core network

Important properties of the core network are high capacity, high scalability and high level of protection. On the contrary, the specific capital cost of the components is of less importance as many users share the cost of the equipment.

The high bandwidth of optical transmission through fibres has triggered the development of optical core nodes able to support the packet switched nature of IP and the Internet in particular. Traditionally, the optical input signal is converted into an electrical stream of bits, which contains the data that should be forwarded to different output ports. This requires electronics operating at the optical bit rate with significant power consumption as consequence. The motivation for optical packet switching is the expected reduction in power and space by enabling the packet switch to let the payload pass transparently without reading each bit, why most of the electronic control is reduced from bit rate to packet rate speed, thus reducing power consumption.

The functional architecture of an optical packet switch is shown in Fig. 2.2 highlighting the main functions. The optical input signal formatted in packets with a header and a payload arrives at the input interface, and part of the signal power is tapped to the electrical control system. Here, the header, which can be of lower bit rate, is read in order to determine the exact arrival time and the requested output port. If the switch fabric operates in synchronous mode, the packets are synchronised by controlling optical fibre delay lines. After switching through the optical switch fabric, a new header is synchronised on to the packet, the quality of the packet is improved through the signal regenerator, and finally any conflict at the output port is solved using, e.g., optical buffers. It is noted that the order of the functions depends on the used technology and the implementation.

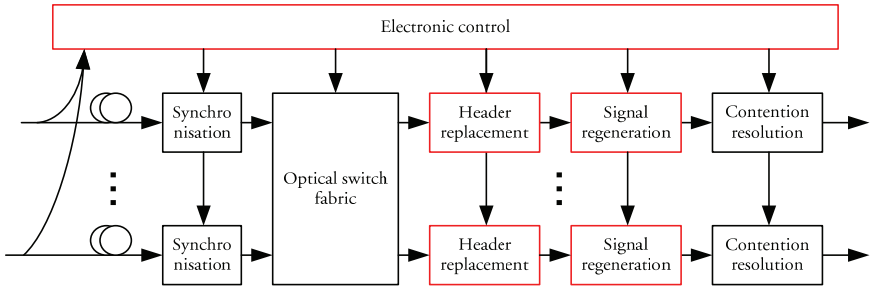


Fig. 2.2: Main blocks in an optical packet switched router. The red blocks are considered further in this thesis

It is obvious that such a setup requires strict electronic control to determine delay and jitter at the input, controlling of the switch fabric, reading and writing the header etc. The realisation and control of two of these functions, namely the signal regeneration and the header management is considered in this thesis in chapter 3 and 4, respectively. Considering the regeneration, the main contribution of this thesis is on the equalisation of packet to packet power variations prior to optical regeneration. For the header management, the difficult resynchronisation of the header to the

packet is avoided by introducing a novel scheme that allows the same header to be reused through the complete optical network domain. This is a prerequisite for any possible improvement of the signal quality, which again is required in order to cascade more than one optical switch.

2.1.2. Access networks

In contrast to the core network the main focus in the access network is on the reduction of component, deployment and maintenance costs. The cost of the equipment is usually only shared by a single or few end users, and the large number of end users requires very streamlined and automatic maintenance procedures. On the other hand, it is usually not catastrophic if the connection to the end user should break down as the fault only has a limited impact on a single or a few users.

Optics in the access networks are introduced by the Fibre to the Premises technologies, where the premises can be the residential or business building, the nearby wire closet or the complete path to the end user terminal. One of the most deployed FTTx technologies is *Ethernet Passive Optical Networks* (EPON), which today is heavily deployed throughout the world with South East Asia as the dominator. In this thesis, however, the focus will be on the transceiver components, optical and electronic, for future high bandwidth solutions with bit rates of 10 Gbit/s.

Hence, in chapter 5 a high capacity point to point FTTH solution is designed and evaluated in order to determine potentially cost efficient solutions and indicate the optimal choice of lasers, receivers etc.

2.1.3. Administrating optical circuits

Apart from controlling optical components, another topic considered in this thesis is the control and management of optical light paths in the core network. The main objective of such work is to increase the utilisation of the bandwidth while providing *Quality of Service* (QoS) guarantees to high demanding applications like video conferencing, high quality distributed video production etc.

Today, the provisioning of circuits (optical or electrical) in the core network requires communication with the network management interface through human interaction. In reality, the setup procedure includes several emails, phone calls and other human interaction, which limits the dynamicity and increases the possibility or faulty configuration. Furthermore, human interaction is not cost-efficient.

In chapter 6, the first steps towards the realisation of a user/application triggered dynamic optical network is considered, which is based on *Generalised MPLS* (GMPLS) and ASON technologies, as will be introduced in sections 2.2.2 and 2.3. The focus in this work is on the vertical integration of the application with a resource provisioning layer that connects to the control plane of the core transport network.

2.2. Multi Protocol Label Switching

Multi Protocol Label Switching (MPLS) is used as control plane framework throughout the thesis, why this section counts as a general overview of MPLS and Generalised MPLS. A reader familiar with these technologies can skip to section 2.3, and further information can be gained from, e.g., [6] and [7].

The connection less nature of the existing *Internet Protocol* (IP) introduce some problems that were not foreseen, when the protocol was developed in 1974 [1]. Firstly, to optimise the utilisation of the relatively small number of IP addresses[1], it has been necessary to flatten the address hierarchy[2], which on the other hand has increased the requirements to the size of the lookup tables in the routers. This is because the bit mask length is variable, and it complicates and increases cost of lookup processing in the routers due to necessary concepts like *Longest Prefix Match* (LPM)[3]. Secondly and more important, IP have limited capabilities of *Traffic Engineering* (TE), i.e., it is not generally possible to determine an explicit path through a network or a network segment, as the routing decision is taken independently in each router[4]. This disables or at least complicates the possibility of offering guarantees on parameters like maximum delay, delay jitter, bandwidth etc.

2.2.1. Label Switching Concept

MPLS addresses these issues by mapping the IP packets to *Label Switched Paths* (LSP), hence introducing a connection oriented transport of the IP data while reducing the router lookup processing. In this section, the map-

[1] The number of usable public IP addresses is around 3.6 billion.

[2] The introduction of *Classless Inter Domain Routing* (CIDR) increased the utilisation of the existing IP addresses by using a flattened network address hierarchy.

[3] If more than one match for an IP address exists, one with 16 bits and one with 24 bits, then the route for the 24 bit match is chosen.

[4] The ToS field in IP provides only diffServ QoS

ping of IP packets to LSPs are reviewed, and the concept of hierarchical MPLS networking is described. While MPLS supports any network layer protocol, only IP is considered in this context due to the numerous applications that are based on IP.

2.2.1.1. Labelling IP packets

MPLS was standardised by *Internet Engineering Task Force* (IETF) in 2000 [8], based on Cisco's Tag Switching and Ipsilon's IP Switching technologies [6]. It is characterised by a separation of the network in edge and transit *Label Switch Routers* (LSR) denoted E-LSRs and T-LSRs, respectively. The E-LSRs are responsible for determining a path through the network, while the T-LSRs are responsible for forwarding packets within the network utilising only local information.

In Fig. 2.3, an IP packet is transported from host A to host B along an LSP within the MPLS network.

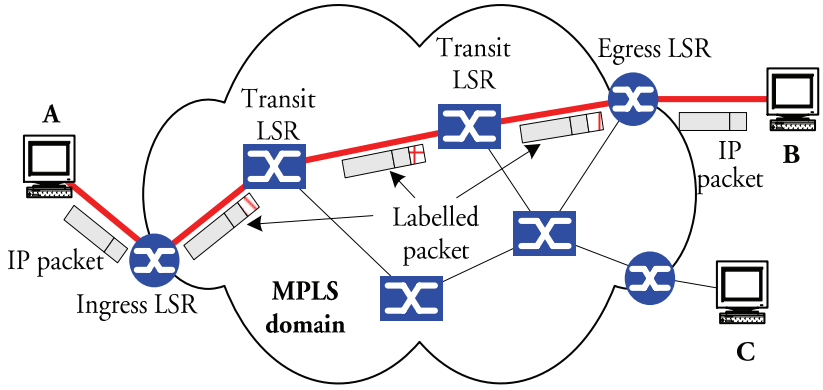


Fig. 2.3: IP packet forwarded from host A to host B in MPLS network

The IP packet is generated at host A and transmitted to the *Ingress LSR* (I-LSR), where it is labelled. The label refers to the LSP from the I-LSR to the *Egress LSR* (E-LSR)[1]. In each of the T-LSRs, the label is read, and a forwarding table lookup is done to determine the output label and the output port. The label has only local significance between two LSRs, and it can be inserted between the network layer and the link layer header as a 32 bit shim header.

In standard MPLS, i.e., when no explicit routing is used, the population of the forwarding table can be thought of as a composition of three sub func-

[1] Note that the abbreviation E-LSR is used for both Edge-LSR and Egress-LSR.

tions. Firstly, all the possible destinations and service classes are partitioned into a finite number of subsets; all packets belonging to a certain subset are treated equally. These subsets are denoted *Forwarding Equivalence Classes* (FEC). Secondly, the next hop for each FEC is determined by the routing component of the LSR. This function can utilise any routing protocol like, e.g., OSPF[1] or RIP[2]. The third sub function is the mapping of labels with the neighbouring LSRs to ensure label integrity of the MPLS network. This function is maintained by a label distribution protocol like the *Label Distribution Protocol* (LDP), *Constraint Based LDP* (CR-LDP) [10] or *Resource Reservation Protocol – Traffic Engineering* (RSVP-TE) [11].

This is illustrated in *Fig. 2.4*, where the first sub function has been used to create the four FECs, A, B, C and D. The second sub function utilises the routing protocol to populate the routing table; a function equivalent to normal IP routing to bind the FEC with output ports.

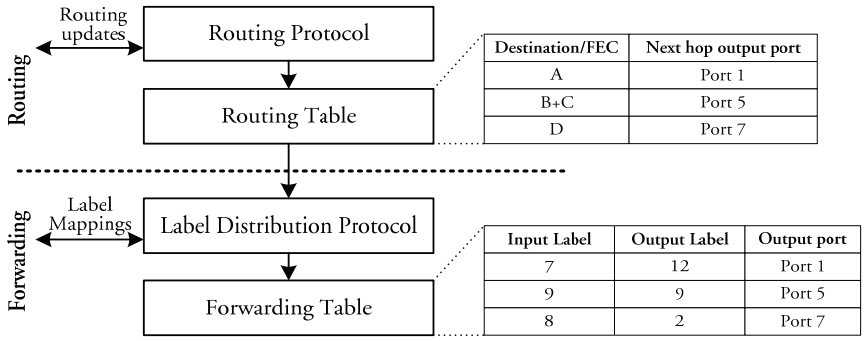


Fig. 2.4: Population of routing and forwarding table in LSR

The FECs are mapped to LSPs through the third sub function, hence separating the routing and the forwarding planes. The result is the forwarding table or *Forwarding Engine* (FE), where the input labels are associated with an output label and an output port. Hence, three sub functions are used for maintaining the forwarding table in each LSR. Only the mapping of label with FECs is special for MPLS compared to IP.

After an LSP has been advertised to the Edge-LSR, it is possible to transmit IP traffic along the LSP by labelling the packets accordingly. The label is included as a shim header or included in the encapsulation.

[1] OSPF: Open Shortest Path First protocol. See, e.g., [9]

[2] RIP: Routing Information Protocol. See, e.g., [9]

2.2.1.2. Label distribution

The distribution of the labels in MPLS networks follows one of the main four label binding schemes.

- Independent downstream unsolicited
- Ordered downstream unsolicited
- Independent downstream on demand
- Ordered downstream on demand

In the downstream unsolicited schemes, the downstream LSR is responsible for advertising labels to its upstream peers, while in the downstream on demand schemes, the upstream LSR is responsible for requesting labels whenever required. In independent binding, a downstream LSR should advertise a label even if it has received no bindings from it own downstream peer. On the contrary, ordered binding requires that a label is received from the downstream LSR (or the LSR is itself an egde-LSR) before advertising to its upstream peers.

In most real scenarios only ordered downstream on demand binding is used, why this is illustrated in Fig. 2.5. The upstream LSR needs a label for a certain LSP, where the next hop is the downstream LSR, why it sends a request message. The downstream LSR forwards this message to the egress LSR, which returns a FEC to label mapping, which again is forwarded to the upstream LSR.

Fig. 2.5: Ordered distribution on demand

Ordered downstream on demand distribution is required for explicit routing enabling traffic engineering as discussed in the following section.

2.2.1.3. Explicit routing

One of the major advantages of MPLS is the support of explicit routing. Here, all the routing functions are moved to the E-LSRs, thus the routing component in the T-LSRs is entirely avoided. Fig. 2.6 shows how an explicit path is established in an MPLS network.

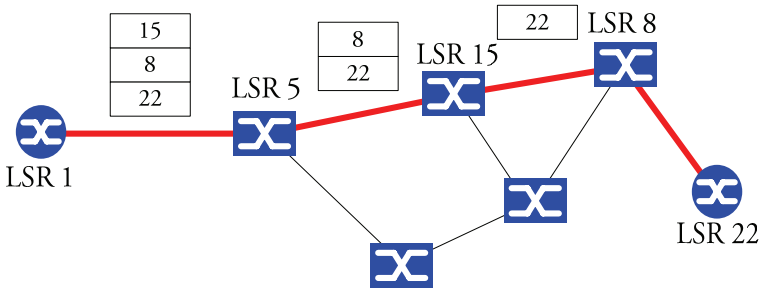


Fig. 2.6: Explicit route establishment. The signalling messages include information on which output to use.

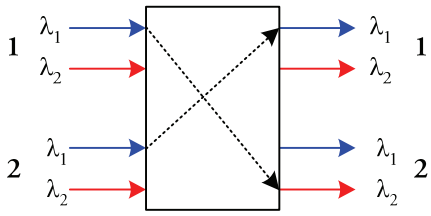
An explicit route from I-LSR 1 to E-LSR 22 through T-LSR 5, 15 and 8 is requested. When the LSP is requested, a stack of all the LSRs en route is included in the request message. From I-LSR 1, the stack contains entries for LSR 15, 8 and 22. At LSR 5, the first entry is popped and the request is sent to this LSR, i.e., in LSR 5, the request is forwarded to LSR 15. A label mapping is returned in the opposite direction, and no routing components are necessary in the T-LSRs. The first LSR 5 is not included as it is implicitly included in the LSP from E-LSR 1.

Two obvious advantages are observed using explicit routing. Firstly, as it is possible to determine the route explicitly, applications like traffic engineering and restoration can be supported, because it is possible to determine path disjoint from a congested shortest path. Secondly, the functional requirements to the T-LSRs are significantly reduced, which is interesting for the development of optical packet switched networks, as the optical components for the T-LSRs offer only limited functionality.

2.2.2. GMPLS and MPλS

Generalised MPLS (GMPLS) extends MPLS from supporting only packet switching to include circuit switching. Hence, GMPLS supports packet switching, time division multiplexing, fibre switching and wavelength switching [12]. The latter is also denoted *Multi-Protocol Lambda Switching* (MPλS), which thus is a subset of GMPLS.

As a simple example, a 2x2 port Optical Cross Connect (OXC) is shown in Fig. 2.7, where each fibre comprises two wavelengths.



Input Label (port, λ)	Output Label (port, λ)
(1,1)	(2,2)
(2,1)	(1,1)

Fig. 2.7: 2x2 MPλS switch supporting 2 wavelengths and the corresponding forwarding table

The granularity of the switching is wavelength based, and each wavelength can be converted to the same or another wavelength on whichever of the output fibres.

In this example, λ_1 on fibre 1 is directed to port 2 with conversion to λ_2, and λ_1 on port 2 is simply directed to λ_1 on output port 1 without wavelength conversion.

It is noted that the forwarding table uses the same format as for standard MPLS. This means that the same control and management plane can be used for controlling GMPLS, MPλS and MPLS networks; the difference is whether the labels are based on port and wavelength or on the header. In the example, the incoming label corresponding to λ_1 on fibre 2 is denoted (2, 1).

Summing up, MPλS is a subset of GMPLS, which use MPLS control plane for circuit switched networks. The label is created from the fibre number, the wavelength number and/or a certain timeslot of the signal.

2.2.3. Hierarchical labelling

MPLS supports a hierarchical networking structure, where an MPLS network can contain several sub MPLS domains. Whenever a packet enters a new domain or sub domain, a label for that domain is added to the packet. Hence the MPLS forwarding information is extended from a single label to a variable depth stack of labels denoted the label stack.

Consider a network hierarchy as shown in Fig. 2.8 with four levels. Two LSPs through the network is pre-established; the red (solid) LSP connects the left and centre access ring through level 0 and 1, while the blue (dashed) LSP connects centre and right ring through all four levels.

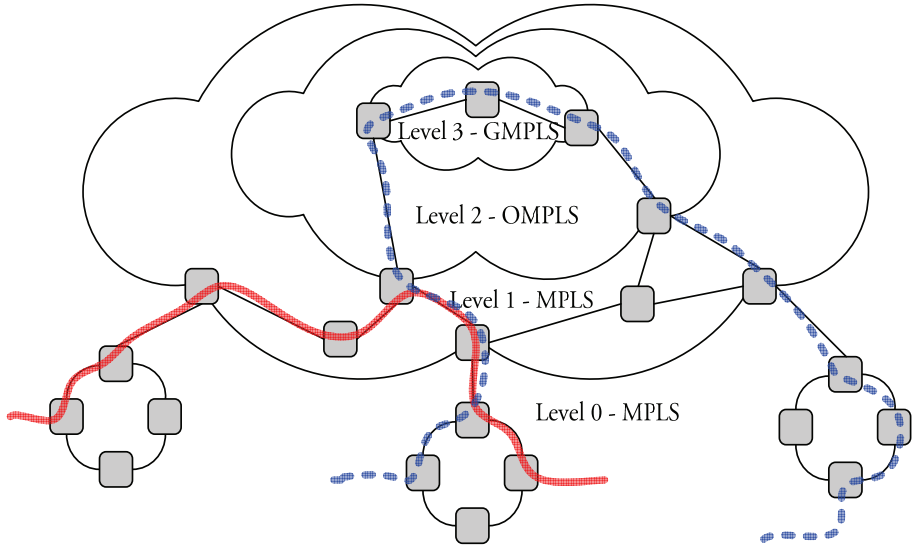


Fig. 2.8: MPLS networking hierarchy

When an IP packet enters the red (solid) LSP (in the access ring), a label is added. Then, upon entering level 1, another label is pushed to the label stack, which now contains two labels. The forwarding within level 1 is based solely on the outer label, which is popped at the exit level 0/1 edge. Four labels are required for a packet travelling through the blue (dashed) LSP.

From level 0, the higher level network is considered as a direct connection between the level 0/1 edges. Hence, some LSRs are considered as T-LSRs for a lower level and E-LSRs for the higher network levels.

The advantages of hierarchical network architecture are scalability and easier integration of different technologies. From the lower levels, the network is logically considered as a major distributed switch not considering the T-LSRs in the higher levels. In the higher levels, several flows utilise the same LSP tunnel why aggregation of the lower level LSPs is possible. This has the potential to increase the utilisation as demonstrated in [13]. Each level can use a different technology; while, e.g., level 0 and 1 uses normal electronic MPLS, level 2 might use optical packet switching (OMPLS). Furthermore, level 3 could be a circuit-switched GMPLS domain.

Hence, logically the entire network is fully label switched although different technologies might be used with adaptation and aggregation between the packet and circuit switched domains.

2.3. ASON and multi domain networks

The (G)MPLS network model, as discussed in the previous sections, is only valid for a single administrative network domain including sub domains. Hence, if one requests a traffic engineered path between host A and host B in Fig. 2.9, multi-domain traffic engineering and path setup has to be considered.

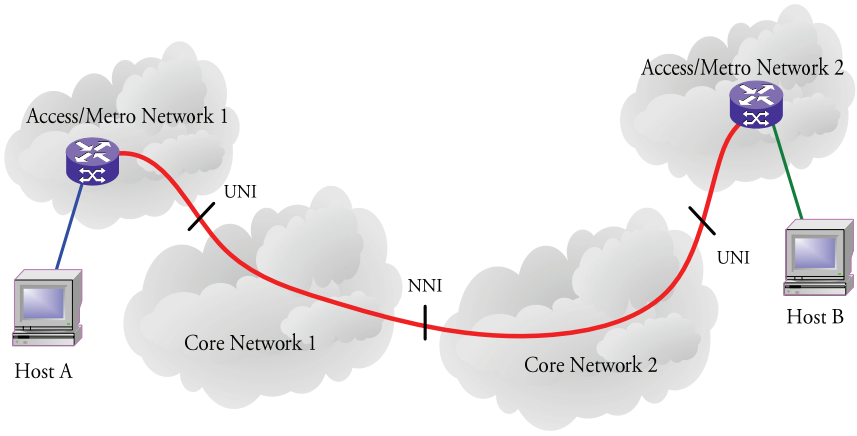


Fig. 2.9: Multi-domain traffic engineering using reference points

The connection from host A to host B passes two access or metro networks in addition to two core networks. GMPLS does not support such multi-domain path setup, as the different networks might belong to different telecom providers, research networks or other providers. These do not share topology information over administrative domains, as required by GMPLS.

2.3.1. Role of standardisation bodies

Within the standardisation body ITU-T[1] the establishment of an interoperable control plane between administrative network domains is considered. Automatically Switched Optical Network (ASON) [14] defines reference architectures for multi domains, the components included and the interfacing between them. In contrast to GMPLS, which is standardised through the IETF, ASON heavily draws on traditions from the telecom concepts like

[1] ITU-T: International Telecommunication Union, Telecommunication standardisation sector of ITU.

SDH[1], SS7[2] and ATM[3] [15]. One objective is to provide a control plane for traditional telecom networks based on SDH transport technologies. As the ASON reference architecture covers a broad area the reader should consult the standard [14] for detailed information.

The IETF, which is the body developing the framework for GMPLS and the related protocols, also considers the integration of administrative domains. The GMPLS mainly considers a peer model where the same "unified" control plane is used for the complete network. In contrast to ITU-T, IETF provides protocols supporting the required functions.

The different solutions proposed by the IETF and ITU-T bodies are sought integrated in the *Optical Internet Forum* (OIF), which contains members from both IETF and ITU-T. Through the *Implementation Agreements* (IA) OIF attempts to integrate the IETF protocols with the ITU-T reference architecture. In the context of integrating multi domains the, UNI IA 1.0 [16] defines the UNI reference points and the usage of the RSVP-TE and CR-LDP signalling protocols for implementation.

2.3.2. Reference points

The interfacing between the network domains is defined by the specification of reference points, which details the level of trust and information passed through the interfaces.

2.3.2.1. User Network Interface

Referring to Fig. 2.9, the *User Network Interface* (UNI) is used to describe the interface between a user and the core network. Here the term "user" can denote a specific user, an application or a client network.

The following functions are supported by the interface description by OIF.

- Resource discovery
- Connection control
- Connection status

It is noted that topology and routing information is not passed through the interface.

---

[1] SDH: Synchronous Digital Hierarchy.

[2] SS7: Signalling System 7

[3] ATM: Asynchronous Transfer Mode

2.3.2.2. Network-Network Interface

Referring to Fig. 2.9, the *Network-Network Interface* (NNI) specifies the reference point between two control domains, e.g., two distinct GMPLS domains. The specification of the NNI is further subdivided into *External NNI* (E-NNI) and *Internal NNI* (I-NNI). The main difference is that routing information is passed through the I-NNI, which denotes interfaces within a single administrative domain, e.g., between vendor specific sub domains.

Apart from differences in the topology flooding, the main functions of the NNI are the following similar to the UNI functions:

- Resource discovery
- Connection control
- Connection status

Using the reference points in the networks, it is thus obvious that I-NNIs can be used within a GMPLS domain, why E-NNIs are used to separate interoperable GMPLS domains.

2.3.3. Connection control

Basically, three types of connection establishment exist depending on, whether an interoperable control plane exists.

Today, the establishment of connections follows the scheme shown in Fig. 2.10.

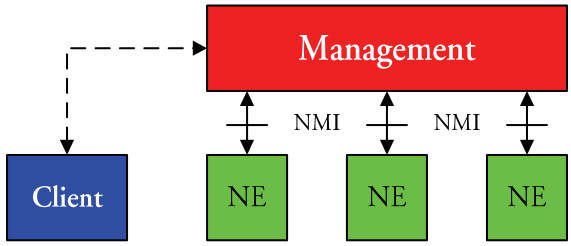


Fig. 2.10: Permanent connection setup. The management is for simplicity here shown as a single entity.

The client requests a connection to the management of the network domains through email, phone, web-forms or whatever. The management plane uses the *Network Management Interface* (NMI) to setup the connec-

tions in each network domains en route. Obviously, the time scale for such connection setup is not in the second scale[1].

The first step towards the realisation of a control plane can be realised with soft-permanent connections as shown in Fig. 2.11.

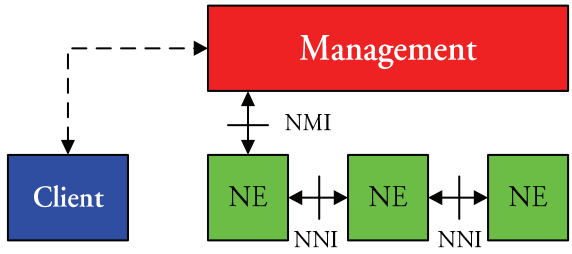


Fig. 2.11: Soft-permanent connection setup

The "UNI" is still based on "out of band" signalling using human interaction. However, only the first network domain is contacted, and further setup is done in the control plane through the E-NNIs. This approach relaxes the requirements to the management plane.

The switched connection paradigm is realised by providing a UNI towards the client as illustrated in Fig. 2.12. Here, the UNI provides authorisation, authentication and other functions that were implicitly done through the network manager.

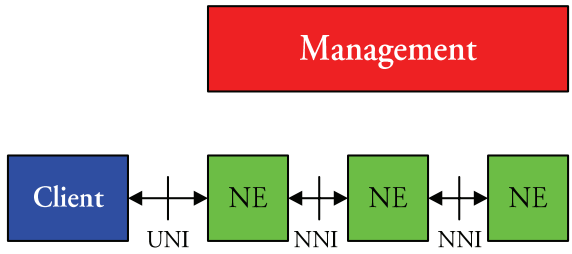


Fig. 2.12: Switched connection setup

Hence, the management plane is not directly involved in the connection establishment and the setup procedure can now be done in second and minute timescales as will be used in chapter 6.

[1] In practical examples time scales in the month scale has been observed.

2.4. Summary

Optical infrastructures are used in both the access and core network and it should be controlled in several aspects.

In this chapter some specific aspects of electronic control of optical infrastructures has been selected for further consideration through the thesis.

In the core network one objective is to increase flexibility and reduce costs per bit by introducing optical packet switched networks. Two challenges within this area have been selected, namely equalisation of packet to packet power variations in combination with optical signal regeneration and schemes for avoiding header modification in the packet switch.

For the access network the component choices of lasers and receivers have selected for further considerations. Here, costs of components are highly important only few users share the costs.

The MPLS networking framework was introduced, which provides traffic engineering capabilities to IP network. While MPLS is designed for packet switched network, GMPLS extends the framework to include circuit switched network, i.e., fibres, circuits and wavelengths. Hence, GMPLS can be used to control optical wavelengths and connections within a network domain from the edge of the network.

While (G)MPLS provides traffic engineering for a single network domain, ASON technologies defines interoperable interfaces between heterogeneous network domains, which is used for multi-domain network management. The concepts from permanent via soft-permanent to switched connections were described.

3. 3R with Power Equalisation

The exploitation of photonics in the transport networks today is mainly used for point-to-point connections, i.e., optical technologies are only used for the transmission part between electrical switch nodes. For optical nodes, however, a number of parameters like power level, *Optical Signal to Noise Ratio* (OSNR), timing jitter and dispersion differ on a channel or packet basis, which has to be taken into account.

Any active switching device reduces the OSNR, and non-ideal suppression of neighbouring channels contributes to crosstalk. Especially interferometric crosstalk, i.e., crosstalk from channels in the same wavelength will significantly reduce the quality of the optical signal through the switch [1]. These reductions of the signal quality make it unfeasible to cascade a large number of switch nodes, which is required to develop flexible all-optical transport network. It is therefore necessary to regenerate the optical signal thoroughly at the optical switch node. This is called 3R regeneration for Reamplification, Reshaping and Retiming.

In addition, for a packet switched network, the packets arrive from different sources with different power level, which leads to a variable power level at the output. This is shown in Fig. 3.1 for a 2x2 switch, where the packets at the output have two different packet-to-packet power levels.

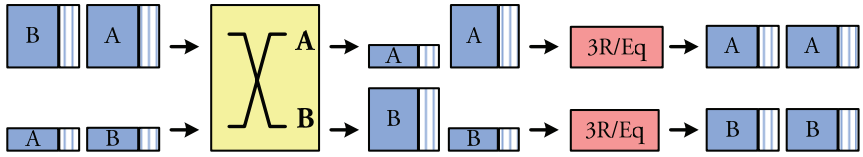


Fig. 3.1: Packets from different sources have different power level. The "3R/Eq"-block accounts for 3R regeneration and power equalisation.

It is therefore necessary within an optical packet switched network to both regenerate the signal *and* equalise the packet-to-packet power level. In the figure, these two functions are illustrated with the "3R/Eq" block.

This chapter addresses the development, implementation and integration of a power equaliser operating in combination with a 3R regenerator for optical packet switched networks. The work is closely related to the European IST project DAVID, why section 3.1 provides a general overview of the

DAVID project. Secondly, in section 3.2 the general concept of regeneration is described, and approaches for electrical and optical 3R regeneration are reviewed. The regenerator proposed within the DAVID project is then detailed in section 3.3 including theoretical and experimental analysis of integration with power equalisation. In section 3.4, the development of the control electronics for the power equalisation is considered followed by experimental results in section 3.5 within an environment with packet-to-packet power variations. Finally, a summary is provided in section 3.6.

The chapter is mainly based on publications [Publ. 8], [Publ. 9] and [Publ. 10].

3.1. European IST project DAVID

The potential benefits and shortcomings of optical packet switching are addressed in the DAVID project, funded by the European Commission. The main objective of the project is to propose and demonstrate a viable approach towards optical packet switched networks. This section provides an overview of the DAVID project with focus on those parts that are relevant seen from an electrical control perspective.

The objectives of the DAVID project are presented in section 3.1.1. This gives an overview of the main purpose of the project and how it differentiates from previous projects in the area. In section 3.1.2, the networking concept that is used in the project is described; the *Metropolitan Area Network* (MAN) and the *Wide Area Network* (WAN) part of the network are presented with focus on switching concepts and the network control. The developed DAVID demonstrator is finally presented in section 3.1.4.

3.1.1. Project objectives

It was stated in the introduction that a migration to optical packet switching and optical technologies in general has the potential to reduce the power and space requirements that are expected to become a bottleneck for electrical switch nodes. The DAVID project aims at suggesting a feasible approach towards OPS. This takes into account not only development of optical switch components, but also the integration with existing networks, traffic performance studies, benchmarking, and control and management issues. Finally, the developed concept is verified through a demonstrator. In all parts of the project, it is the intention to use a pragmatic approach utilising today's and tomorrow's advantages of both the electronic and the optical technologies.

To obtain these goals, the DAVID consortium spans a broad range of partners from academic institutions over industrial partners to network operators. This adds up to fourteen partners from nine European countries [4].

The feasibility of optical networking has been investigated and evaluated in a number of previous projects. The ACTS[1] KEOPS[2] project focused on the optical implementation of the processing functionalities like synchronisation, regeneration and switching [17][18]. The British WASPNET[3] and the OPERA[4] project suggested a switching approach utilising wave-guide routing based on *Arrayed Waveguide Grating Routers* (AWGR) and *Semiconductor Optical Amplifiers* (SOA) for wavelength conversion [19][20]. Another project, HORNET[5], considered the use of OPS in a 2.5 Gbit/s MAN network [21]. The results of these and other projects are heavily used within DAVID to make descisions regarding contention resolution, packet format, switching approach etc.

As different requirements apply to the access and the core network, the approach in DAVID is a separation of the network in a MAN and a WAN part. The MAN part comprises a ring structure, whereas the WAN comprises a meshed structure. The ring structures are easier to control in expense of an inferior scalability. From a management perspective, it is important to have a unified management plane regardless whether a network segment is within the MAN or the WAN. This is one of the reasons for considering the DAVID network as an MPLS based network; it furthermore has an impact of the traffic studies and the requirements to the physical implementation.

The proof of the concept is of very high priority. Therefore, a considerable part of the efforts is directed towards the development of a demonstrator for physical verification of the proposed concepts.

3.1.2. Networking concepts

Different requirements apply to different parts of a network. In the MAN or access network, it is important to reduce cost, as relatively few clients share the expenses. Therefore, focus is on cost-efficient technologies avoid-

[1] ACTS: Advanced Communication Technologies and Services

[2] KEOPS: Keys to Optical Packet Switching

[3] WASPNET: Wavelength Switched Packet Network

[4] OPERA: Optical Packet Experimental Routing Architecture

[5] HORNET: Hybrid Optoelectronic Ring Network

ing complex buffering etc. In the WAN or core network, the main issues are capacity and scalability, why technologies for increasing capacity while keeping power consumption to a minimum should be considered.

The network architecture in DAVID reflects these dissimilar requirements, although both parts of the network use fixed length packets with synchronous operation [22]. In Fig. 3.2, the DAVID network architecture is shown with the separation in the MAN and the WAN part.

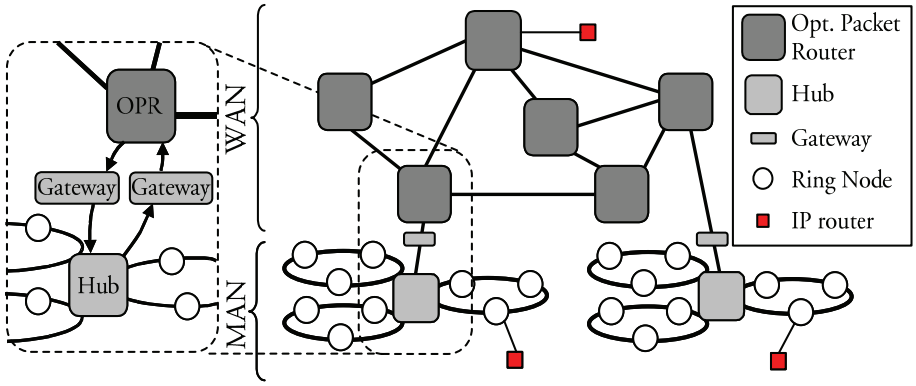


Fig. 3.2: DAVID network architecture

The DAVID MAN comprises a number of rings; these are used to collect and distribute traffic to the clients through the ring nodes. The hub forwards the optical packets between the rings and through the gateway to the WAN. Hence, logically the connection to the WAN is considered as an extra ring, and the gateway is responsible to solve contention between packets flowing from MAN and WAN and for aggregating packets for the packet format used within the WAN [22].

The WAN is a meshed network connected by *Optical Packet Routers* (OPR), as will be described later.

Referring to Fig. 2.8 it is observed that the WAN is well suited for use with MPLS and GMPLS. Hence, it is possible to ensure scalability and support for QoS by aggregation of traffic in lower order LSP tunnels [22].

3.1.2.1. MAN

The Metropolitan Area Network (MAN) is a set of ring networks connected by an optical hub as shown in Fig. 3.2. Each ring comprises 32 data channels on 32 wavelengths.

The objective of the hub is to interconnect the rings in the MAN and to provide access to the WAN through the gateway. The optical hub is bufferless, and packet loss is thus only avoidable, if full control of the rings is ensured. Therefore, a *Medium Access Control* (MAC) protocol is developed to control and manage the resources of the rings. This requires that a control channel is added to each ring as a 33rd wavelength, and this channel distributes information of the destination of each timeslot on each wavelength.

Traffic from client IP routers and other legacy networks are thus transmitted to the ring in time-slots depending on whether the destination is on the local ring or whether the data is destined for another ring in the MAN.

3.1.2.2. Ring Node

Basically, the ring node serves as gateway between the client layer networks and the DAVID optical packet MAN. In Fig. 3.3, the architecture of a generic ring node is shown. The slotted optical data pass through the *Optical Packet Add and Drop Multiplexer* (OPADM), while the control channel is dropped to the ring node. If a packet is pending for the actual ring pair or if a packet is to be received, the ring node then updates the control channel. Neither the data packets nor the control packets are bit synchronous within the fixed length timeslot, why *Burst Mode Transceivers* (BMT) are required for transmission and reception of the optical signals.

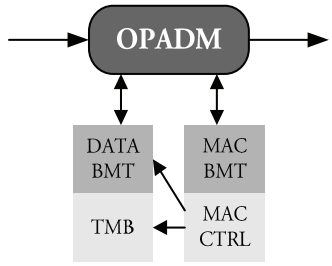


Fig. 3.3: Generic Ring Node.

The *Traffic Manager Board* (TMB) is responsible for conversion, i.e., segregation, aggregation and encapsulation, of the packets to and from the client layer networks. This includes queuing of the packets and shaping the traffic. A very important functionality is the aggregation of variable size packets to the fixed length optical cells. Studies within DAVID have addressed this issue to identify the optimal parameters for aggregating the client layer packets without introducing an unacceptable delay [23][24].

3.1.2.3. Traffic Shaper

When data is to be transported from the MAN ring to the legacy network, the bandwidth is significantly reduced to two *Gigabit Ethernet* (GbE) interfaces. Hence, traffic shaping is necessary to cope with bursts exceeding the bandwidth limitations of the GbE.

Although the traffic shaper function in the TMB is a classical scheduler problem, special requirements apply due to the bit-rate and the general structure of the TMB. Therefore, a novel scheme, *Provision based Fair Scheduling* (PFS), has been developed, which uses a provision mechanism to estimate the length of packets to be scheduled. This subsection describes the novel scheme based on [Publ. 11].

The traffic from the MAN is divided into a number of queues depending on the origin of the packets, i.e., which node is the source. Hence, the scheduler basically determines which of the queues that is granted access for transmission to the GbE interface. The main objectives are fast and fair access to the output, however, different queues can have different priorities.

The simplest scheduling algorithm, *Round Robin* (RR), operates by visiting active[1] queues one by one transmitting one packet per visit. While this is suitable for fixed length packets, it does not ensure any fairness for variable length packets as is the case at this point in the ring node. Therefore, different scheduling algorithms are evaluated, and it is chosen to base the implementation on a modified *Surplus Round Robin* (SRR)[2] scheduling algorithm [25].

SRR is an extension to RR, where a surplus counter is associated to each queue. Every time a packet is transmitted from the queue, its surplus counter is decremented according to the length of the packet. If the counter becomes zero or negative, the scheduler proceeds to the next active queue. For each full round the surplus counters are updated with a predefined quantum.

Operation of SRR requires that information of the length of a scheduled packet should be known before the scheduler can proceed, i.e., it must determine whether to serve more packets from the same queue or to proceed to the next active queue. This information is in the TMB not always available because it is not known when the memory will be refreshed causing a

[1] Active queues denote queues with pending packets.

[2] Surplus Round Robin is in itself a modification to Deficit Round Robin with the modification that negative deficit counters are allowed, which allow transmission of an unknown length packet.

substantial delay. Therefore, it is necessary to modify the SRR algorithm allowing it to continue although the length information is unavailable.

Provision-based Fair Scheduling (PFS) uses an estimation of the packet length instead of the actual length for decrementing the surplus counters. The estimation can be based on the average packet length, the length of the previous packet of the flow or any other value. Obviously, this does not ensure fairness, as the estimated length may not exactly match the actual length. Therefore, when the packet is finally transmitted the surplus counter is updated with the difference between the estimated and the real packet length. Hence, the long term fairness is ensured.

In Fig. 3.4(a), PFS is used to schedule between three queues. The provision is based on the average packet length of two and the maximum packet[1] of length four is scheduled for each active queue, i.e., two packets are scheduled from each queue in the initial round which is A, B, D, E, H and I. Assuming that the actual packet length has become available before next round, the surplus counters are updated. Thus the counters after the roundly update for the three queues equal 3, 6 and 0, respectively. In the next round, an extra packet is scheduled from the second queue and no packets are scheduled from the last queue.

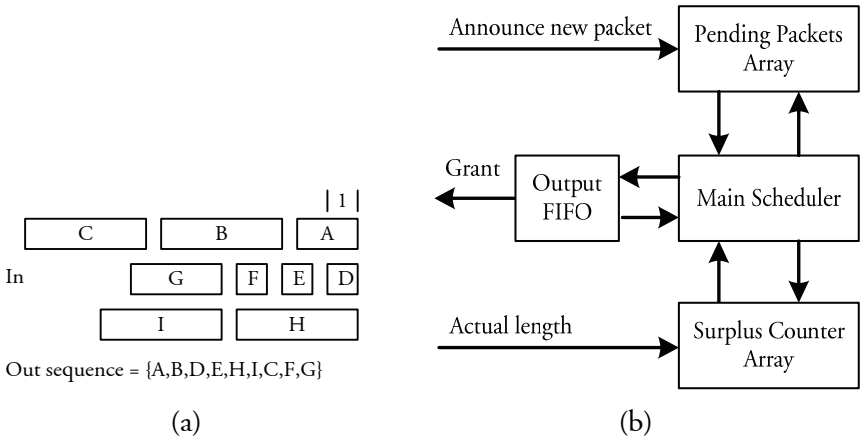


Fig. 3.4: Scheduling with PFS (a) and architecture for implementation (b).

The implementation of PFS in *Field Programmable Gate Array* (FPGA) is sketched in Fig. 3.4(b). The four main blocks operate independently, which allows fast and simple design.

[1] The "maximum length" indicates the maximum length of packets supported for the network.

The pending packets array contains a counter for each queue. Whenever a packet arrives to the traffic shaper, its corresponding queue is updated, while the packet is stored. The main scheduler decides which queue should have access to the output based on the surplus counters and the active queues defined by the surplus counter array and the pending packets array, respectively. Every time the main scheduler grants access to a queue, the queue number is stored in the output *First In First Out* (FIFO), which is read by the memory management system. The adjustment of the surplus counters is done by signalling the actual length information when the packet is transmitted.

The depth of the output FIFO should be chosen carefully. A small depth FIFO corresponds to a smaller acceptable delay from the management system, while a very deep FIFO reduces short term fairness.

3.1.3. WAN

The architecture of the Wide Area Network has an MPLS based structure, which is utilised to guarantee QoS through traffic engineering and scalability through aggregation and tunnelling [26]. These issues were discussed in section 2.2. Hence, the structure of the WAN, as shown in Fig. 3.2, can logically be structured similar to the hierarchy of MPLS networks in Fig. 2.8. Although the technologies of the switch nodes are different in each layer, the structure provides a uniform network control plane.

In optical packet switched networks the wavelength dimension is potentially used for two purposes. It can be used for routing in a circuit-switched manner, where traffic at the edge of an MPλS segment is aggregated and tunnelled based on the wavelength. Alternatively, the wavelength dimension is used for contention resolution to reduce the number of fibre delay lines as discussed in the next section. The advantages of both are achieved by partitioning the wavelengths into wavebands. Within each waveband the wavelengths are used for contention resolution and the entire wavebands are used for MPλS switching.

The WAN is based on high capacity *optical packet routers*[1] (OPR) or switches. This introduces several technological challenges with the migration of necessary functions from the electronic to the optical domain, as will be discussed in the following subsection.

[1] The names ”optical packet switch” and ”optical packet router” are used indiscriminately within DAVID.

The structure of the WAN is a meshed network, and therefore it is not possible to use a MAC protocol with an out of band control channel. Thus, the packet header is carried in-band, i.e., the header is inserted before the payload in the time domain at the same wavelength as the payload. The length or duration of the payload is not necessarily similar to that of the MAN; however, fixed length packet duration of approximately 1 μs is chosen within DAVID for an intended bit rate of 10 Gbit/s.

3.1.3.1. Optical Packet Switch

The functionality of the optical packet router is very similar to the hub, except that buffers are used instead of a MAC protocol to do contention resolution. During transmission through the optical packet router the payload of the data resides in the optical domain, while the header is processed electronically. The optical switch should include a number of functions: Firstly, the incoming data is demultiplexed, and the packets are synchronised making it possible for the OPR to operate in a slotted mode. Next, the header is extracted for electronic header processing, and thereafter the payload is switched through the optical switch without any conversion to the electronic domain.

Contention resolution is a major issue in optical packet switches as no useful optical RAM has been invented. In the DAVID OPR, this problem is addressed by using a combination of fibre delay lines and the wavelength dimension. The latter significantly reduces the requirements to the number of fibre delay lines as discussed in [27] and [28] in relation to the KEOPS project. The few stages of fibre delay lines that are anyway required are in the DAVID concept implemented as recirculating buffers.

The switch block of the DAVID OPR is shown in Fig. 3.5, where functions like synchronisation, header extraction and insertion are omitted. The demultiplexed and synchronised data streams are regenerated in the 3R block before multiplexing and amplification. The output is chosen through the two-stage space and wavelength selection block; the first stage selects a group of wavelength from the desired input port and the second stage selects the wavelength by choosing which pass-band filter of the multiplexer that should be used. The packets are then regenerated and wavelength converted in the 3R block. Finally, a new header is inserted.

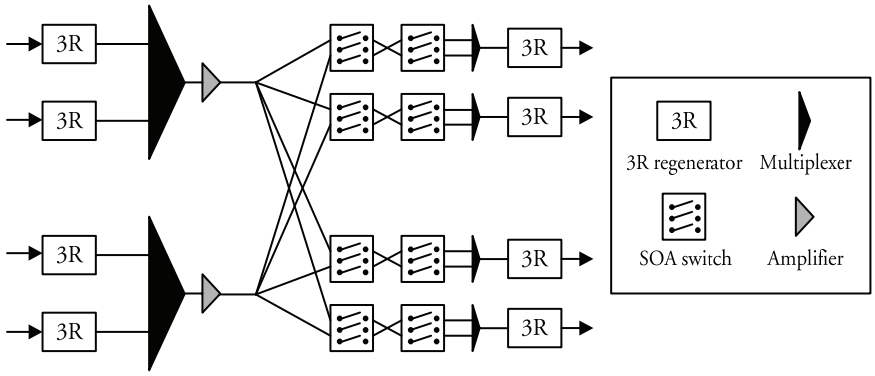


Fig. 3.5: Structure of the OPR

In the packet switch, several sources for signal degradation exist. Therefore, it is necessary to regenerate the signal at the input and at the output. The regeneration of the signal increases the quality of the pulse shape and at the output packets from different sources should have the same power level. Hence optical packet to packet power equalisation is required.

3.1.3.2. Technologies and challenges

The components to build the DAVID network is far from "off the shelf" components. It is thus necessary to evaluate and study the optical and electronic technologies most suitable within the project.

Considering the OPR, several requirements apply to the basic building block, the optical switch elements. Firstly, they should be very fast, which is in the nanosecond time scale, as they should be reconfigured for each timeslot. Secondly, the input interface should have very low polarisation[1] sensitivity, as it is not possible to control the polarisation state of the incoming fibre. Then they should have high on-off ratio to sufficiently suppress the signal in "off"-mode, thus reducing interferometric crosstalk[2]. These requirements are potentially satisfied by the SOA technology. Furthermore SOAs can be integrated in gate arrays [30], which is very important for switching applications.

---

[1] The polarisation of light fluctuates with time and space. The polarisation sensitivity refers dependent a component is towards variations in the polarisation state. Refer to [1] for further information.

[2] Interferometric crosstalk, which is on the same wavelength, is significantly worse than power-addition crosstalk, which is on a different wavelength. Refer to [29] for further information.

The 3R regeneration in DAVID is based on all active SOA based *Mach Zender Interferometers* (MZI), which has been used for regeneration of continuous signals up to 40 Gbit/s [31].

Optical header modification is only slightly addressed within DAVID although it is a necessary function for a label swapping scheme like MPLS. However, different schemes for simplifying this process is studied partly within DAVID as will be described in chapter 4.

Before packets arrive at the OPR they are synchronised such that packets arriving at each input port arrive synchronously. The packets should arrive packet synchronous but not necessarily bit-synchronous.

The optical synchroniser in DAVID comprises two stages. The first stage, the Transmission Wander Compensator, operates on all wavelengths simultaneously. This compensates for variations in delay mainly due to slow thermal variations of the transmission fibre. The second stage, the Jitter Extractor, compensates for small packet-to-packet delay jitter. This jitter may be different for each wavelength, why the fast compensation is done for each wavelength after demultiplexing. Both synchronisers operate by switching the signal through different lengths of fibre delays.

The technological challenges do not only apply to the optical devices and the optical router. The required electronics for controlling the switch matrix, the 3R regenerator, the synchroniser and the header processing should be sufficiently fast for reconfigurations within the guard band of the optical packet. Furthermore, especially for the receivers in the MAN and the header reception, the burst mode transceivers are very difficult to implement; for this several high speed technologies like Galium Arsenide (GaAs) HBT[1], Indium Phosphide (InP) and Silicium Germanium (SiGe) HBT have been investigated. It was decided to use a SiGe process as this offers the best properties regarding performance, cost and integration with CMOS in SiGe BiCMOS.

3.1.4. DAVID demonstrator

The objectives of the DAVID demonstrator are to serve as proof that the DAVID networking concept is technologically feasible and for gaining knowledge of the implementation and the integration of the required technologies.

[1] HBT: Hetero Bipolar Transistor.

The importance of the demonstrator was indicated by the following statement from the feedback of the third technical review of the DAVID project in January 2003.

"The most valuable aspect of the project is the demonstrator, which is unique and will highlight critical aspects in the implementation of optical packet switching"

The physical implementation and integration of the demonstrator was done in the laboratories at Alcatel in Marcoussis, France. The demonstrator is shown in Fig. 3.6 as it looked a few minutes before the final project review.

Fig. 3.6: DAVID demonstrator in Marcoussis

The demonstrator includes a sub-equipped multi-ring optical packet metro network and a sub-equipped optical backbone including the key functionalities to build an all-optical packet switched network. It basically comprises four sub-demos; two for the MAN and two for the WAN. The first MAN demo demonstrates the operation of the ring nodes and the MAC protocol by transmission of real time video applications. The second MAN demonstration shows the operation of the optical hub, which in many aspects is comparable with the optical packet router in the WAN. The two WAN subdemos are showing the viability of optical packet synchronisation and optical regeneration in an environment with fluctuating packet-to-packet power variations.

The four sub-demonstrations were presented during the final project review in October 2003, and during the immediate oral feedback the demonstrator was characterised as "*... an outstanding piece of hardware ...*".

3.2. Signal regeneration

Regeneration is the process of improving an inferior pulse quality; more specifically, 3R regeneration is an acronym for Reamplication, Reshaping and Retiming of the pulse train. This is illustrated in Fig. 3.7, where the vertical dashed lines correspond to the reference time slots of the switch node.

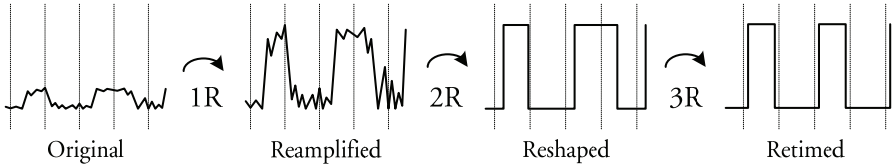


Fig. 3.7: Regeneration of signal through Reamplication, Reshaping and Retiming.

Initially, the signal is of both low power and with significant noise contributions. The reamplification (1R) is simply an amplification of the signal to a reference power level. Usually, an EDFA in a transmission system acts as 1R regenerator, as it amplifies the signal while preserving the pulse shape.

The objective of the reshaping (2R) is to increase the extinction ratio[1] and reduce the noise contributions. This requires a non-linear transfer function as close to a step function as possible. This is sketched in Fig. 3.8, where an ideal transfer function is used, and the inferior incoming signal is improved.

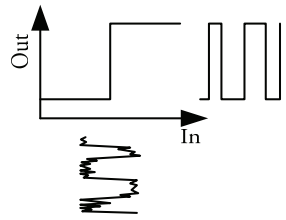


Fig. 3.8: Ideal transfer-function for reshaping.

The ideal reshaping element is simply operating as a decision element, where the flat parts are responsible for suppressing the noise.

The retiming (3R) of the signal is obtained by extracting the clock signal from the incoming pulse train. Then, the clock is used to retime the signal.

In the following subsections, the drawbacks and advantages of electrical and optical 3R regeneration are shortly discussed.

---

[1] Extinction ratio: Ratio between power in a mark and the power in a zero of a pulse

3.2.1. Electrical regeneration

The general approach for electrical regeneration of optical signals includes full reception and retransmission of the signal, which require a broad range of optoelectronic and electronic components. The signal to be regenerated is converted to the electrical domain through a photo-diode, amplified through a chain of amplifiers, and then the clock is extracted and used to retime the pulse through a decision element that is also responsible for the reshaping function. Finally, the regenerated signal is reconverted to the optical domain through a directly or externally modulated laser.

This approach has several advantages. The electronic reshaping and retiming is quite good allowing for an excellent extinction ratio at the output. Then the receiver sensitivity for at 10 Gbit/s is usually in the range of -18 dBm with a standard pin-photodiode and as low as -26 dBm for an *Avalanche Photo Diode* (APD). The latter, however, increase the requirements to the SNR and the electronic amplification. Other advantages are independence towards the polarisation and wavelength of the incoming signal and the possibility of using *Forward Error Correction* (FEC).

The drawback of the electrical solution is bit-rate dependent power consumption, i.e., the power consumption increase significantly with the bit rate due to the electrical amplifiers, the decision element and the laser modulator. Evaluations on the power consumption of electrical regenerators and electrical switch fabric are given in [32]. Furthermore, the electrical technologies usually suffer from speed limitation when the bit rate exceeds 40 Gbit/s.

3.2.2. Optical regeneration approaches

The power and space consumption, in addition to technology dependent bit rate constraints for electrical regeneration, raise the interests for alternative solutions, as previously described. Optical regeneration is potentially less power consuming with increased bit rate, as the power usage tends to be less dependent on the bit rate of the signal to be regenerated. This makes optical 3R regeneration attractive for very high speed all-optical data networks.

A review of issues for semiconductor based all-optical regeneration is given in [33] with focus on the performance of the MZI structure. As examples, 3R regeneration at 40 Gbit/s was demonstrated in 1999 utilising an SOA based MZI structure [34], and the cascade of 100 optical regenerators at 40 Gbit/s was shown in 2001 [35]. In both of these achievements, the SOA based MZI was operated with the differential scheme, i.e., the data signal is

launched to both interferometer arms, with a variation in the amplitude and delay. This approach compensates for limitations in the speed of the SOA due to carrier dynamics. A related component, the *electro-absorption* (EA) modulator has also been used to demonstrate 3R regeneration at 40 Gbit/s through a cascade of two wavelength converters [36]. Although the carrier dynamics of EA-modulators potentially allows for faster regeneration, the interest for this approach is reduced as the necessary optical input power is extremely high (~15-19 dBm).

Other schemes for 3R regeneration include fibre based systems utilising self phase modulation. This is used in [37] to demonstrate 3R regeneration at 40 Gbit/s with highly non-linear fibres for the reshaping function; a fibre loop of one million kilometres including 2500 regenerators in cascade are successfully demonstrated. A related setup is used in [38] to demonstrate regeneration at 160 Gbit/s. Although the fibre based 3R regeneration schemes potentially allows for a faster regeneration, the use of fibre within the regenerator makes them quite bulky. Furthermore, unless precautions are made, phase noise or data pulse jitter is transferred into amplitude noise, which strongly degrades the performance of the regenerator.

Retiming the signal requires that the clock is recovered from the incoming pulse train and then used to remove timing jitter from the signal. This function is usually obtained in the electrical domain after O/E-conversion. While the electrical approaches for clock recovery are well known, they suffer from speed limitations and high power consumption, when the clock rate exceeds 40-80 GHz. The speed limitations of state of the art clock recovery circuits are typically around 40 Gbit/s as demonstrated in [39] and [40].

Optical clock recovery has been successfully demonstrated utilising self-pulsating lasers [41]. These are basically three section *Distributed Feedback* (DFB) lasers, where the injected pulses as well as the self-pulsation modulate the carrier density. According to [41], the self-pulsating frequency can be adjusted between 5 and 22 GHz by altering the bias current to the laser. For continuous data streams, optical clock recovery with self-pulsating lasers has been used to demonstrate clock recovery and 3R regeneration with an ultra long SOA (UL-SOA) at 40 Gbit/s [42]. In [43], self-pulsating lasers for clock recovery has been utilised to regenerate long sequences of clock signals followed periodically by long sequences (400 ns) of zeroes at 10 Gbit/s. This experiment showed excellent recovered clock with output OSNR of 34.6 dB/0.1 nm. A related demonstration in [44] showed successful clock recovery at 10 Gbit/s for real asynchronous packets with a guard band of 25.6 ns and a slot duration of 1 μs. These results documents the capabilities of optical retiming at 10 Gbit/s in packet switched networks with the po-

tential for bit rates exceeding 40 Gbit/s with only a slight increase in the power consumption.

In the literature, all the necessary functions for all-optical 3R regeneration have been documented, and although the optical technologies are less mature than their electronic counterparts, both technologies are viable depending on the application. Thus, the discussion should not be considered as a contribution to a fruitless war for and against optical regeneration. However, the potential low power consumption and the integration with the optical switch core makes optical 3R regeneration very attractive for the DAVID project.

In the DAVID regenerator, which are considered in the following section, the effect of a variation in the packet-to-packet power level is added, and approaches for coping with this are analysed.

3.3. DAVID Regenerator

The regenerator within the DAVID framework differs from most of the published optical 3R regeneration schemes, as it should operate within a packet switched environment with packet to packet power variations. Therefore, it has been considered necessary to incorporate power equalisation control electronics into the regenerator.

The main focus in this section is on two different approaches for controlling the gain in the optical components to compensate for the power variations at the input. These approaches are thoroughly evaluated to identify their advantages and drawbacks. First, however, the specifications to the regenerator are described, and the basic architecture is presented. The two approaches are controlling a preamplifier in a MZI and controlling the gain in a XGM in front of the regenerator. Both approaches are evaluated in detail with manual control instead of control electronics. Although the method in this section does not take into account the fast packet-to-packet power variation, it allows us to compare these two approaches without influence from the control electronics

3.3.1. Regeneration specifications

The 3R regenerator in the DAVID optical packet router is located at the output as shown in Fig. 3.9, and it acts as both 3R regenerator and wavelength converter. The latter is required for conversion between the internal wavelengths for routing in the switch and the external wavelength for transmission and contention resolution.

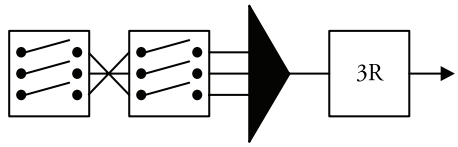


Fig. 3.9: The 3R regenerator is located after the space and wavelength selector in the optical packet router. See also Fig. 3.5.

Because the optical packet router is based on SOA technology for switching, it was chosen to use an SOA based 3R regenerator. This further allows for a compact and robust design compared to fibre based regenerators.

The packet format and expected difference in power level is shown in Fig. 3.10. The packet slot time is approximately 1 µs with a guard band between consecutive packets of around 51 ns.

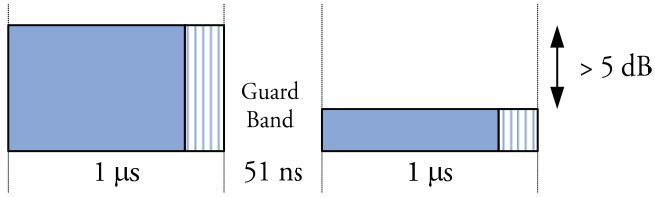


Fig. 3.10: Power variations up to 5 dB between DAVID packets are specified.

The packet comprise after the guard band a preamble of 128 bits 1010 pattern, 128 bits *pseudo random bit sequence* (PRBS) data[1] and 192 bits clock signal (00110011 pattern) before the 8256 bits payload. During the measurements, the latter contained PRBS data as shown in Fig. 3.11.

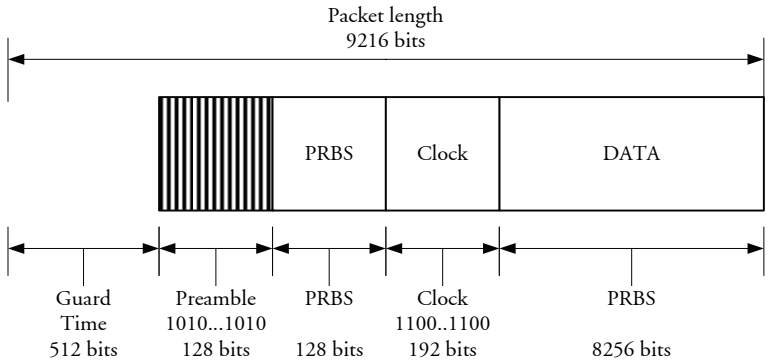


Fig. 3.11: DAVID packet format.

[1] The PRBS sequence is used to model client data.

The different parts of the preamble, e.g., the clock signal is included to cope with constraints in the electrical burst mode receivers in the gateway between the WAN and the MAN. The bit rate of the packets is 10 Gbit/s, but it is intended to ensure operation of the concept at 40 Gbit/s, why the electrical control system should allow for this bit-rate.

The pulse format before the 3R regenerator is *Return to Zero* (RZ) with an optical signal to noise ratio of at least 24 dB/0.1 nm, and with a variation in the packet-to-packet power level of up to 5 dB. The polarisation of the incoming data varies from packet to packet, why the overall polarisation dependency should be as low as possible.

Hence, it is required to do 3R regeneration with fast packet-to-packet power equalisation within the short guard band of 50 ns. Furthermore, the power equalisation should not compensate for long sequences of zeroes.

In the following section, the adopted architecture for the 3R regenerator is explained.

3.3.2. 3R regeneration with MZI stages

The regenerator used in the DAVID project is based on a single SOA followed by one or two MZI devices as sketched in Fig. 3.12.

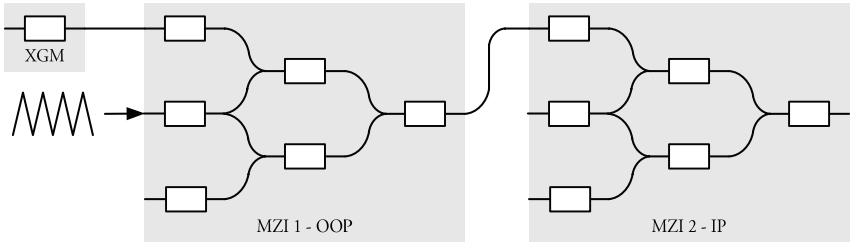


Fig. 3.12: 3R regeneration through XGM and MZI stages. Retiming is obtained by a clocked probe to the first MZI stage. Optical filters between sections are not shown.

The single SOA operates in gain saturation regime as a XGM and serves several purposes. First, it increases the power margins of the input interface as has been exploited since long [45]; in fact up to 8 dB variance are tolerated with two consecutive SOAs [46], and a simple control system are utilised in [47] to allow for even higher variances. It is, however, noted that none of these schemes differentiates between a guard band and a long sequence of zeroes. Secondly, the XGM fixes the polarisation, which is beneficial as the MZIs are very polarisation dependent.

The first MZI stage is operating *out of phase* (OOP) to compensate for the OOP operation of the XGM. The stage is responsible for reshaping and retiming of the signal. The latter is achieved by using a clocked probe signal. The second MZI stage, operating *in phase* (IP), is optionally included to further increase the extinction ratio of the regenerated signal.

The reshaping of the signal through the three stages is determined by their transfer functions, which are sketched in Fig. 3.13. First, in (a) the effect of the gain saturation in the XGM is sketched followed by the transfer function between incoming signal power and outgoing converted power for the MZIs operating in OOP and IP in (b) and (c), respectively.

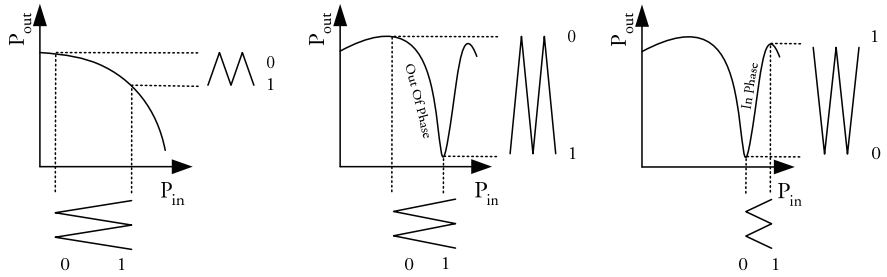


Fig. 3.13: Transfer function for XGM (a) and MZI stages in Out Of Phase (b) and In Phase (c) operation.

It is seen from the transfer function for the XGM (a) that the extinction ratio is inherently reduced due to gain saturation. On the contrary, the *cross phase modulation* (XPM) in (b) and (c) increases the extinction ratio. Note that the gain of an IP conversion is slightly lower than for the OOP conversion. This is caused by gain saturation in the SOAs in the interferometer arms in the MZI device.

In the remaining part of this chapter, only a single MZI stage regenerator will be considered, as this has proven to be largely sufficient. E.g, the combination of the SOA-XGM and the SOA-MZI has allowed the demonstration of 43 Gbit/s FEC compatible RZ transmission with regenerative capabilities in WDM environment (5 channels with channel spacing of 200 GHz) over 30,000 km [48].

3.3.3. Controlling input of MZI

In this section, the first of two approaches for using electronic control to reinforce the power dynamic range is evaluated.

Part of the incoming optical signal is tapped to the electronic control system, which controls the current to the SOA in the input section of the MZI as conceptually sketched in Fig. 3.14. The result is packets with equalised power level before the interferometer arm.

For continuous PRBS data, this scheme was assessed to identify the transfer function from the optical input power to the SOA current for "good" operation, i.e., the most optimal eye-diagram.

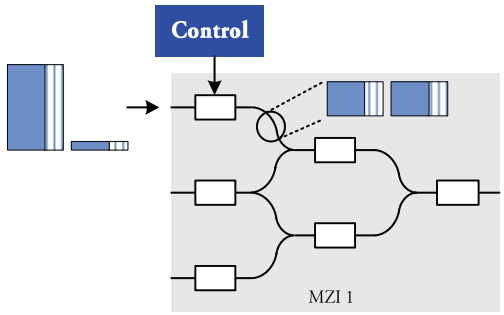


Fig. 3.14: The gain of the input SOA is controlled via the bias current. This ensures equalised packets before the interferometer arm.

Because of the switch structure in DAVID, only up-conversion[1] was evaluated with a packaged MZI interferometer. The effect of wavelength conversion and regeneration is shown in the example in Fig. 3.15, covering conversion from 1540 nm to 1560 nm for different optical input powers. It is noted that a *continuous wave* (CW) signal is used as probe instead of a clocked probe, as it was shown in Fig. 3.12.

The eye-diagrams are open and clear for the full power range of 15 dB. However, a *Bit Error Rate* (BER) measurement (not shown) indicates that the optimal operation range is from -10 to 0 dBm. Measurements for conversion from and to other relevant wavelengths give similar results.

[1] Up-conversion: Wavelength conversion from shorter to longer wavelengths

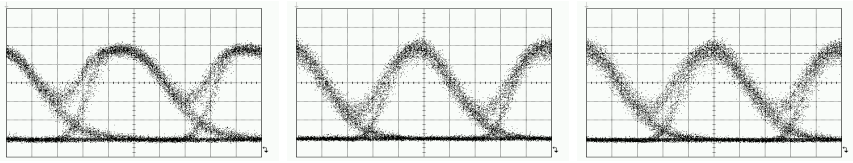


Fig. 3.15: Example of conversion of RZ signal from 1540 to 1560 nm with different optical input powers. A CW source is used instead of a clocked probe signal. The time scale is 20 ps/div and the same values are used horizontally for all three eye-diagrams.

In order to maintain the clear eye-diagrams for the different input powers, the bias current to the input section of the MZI was manually adjusted. The result is shown in Fig. 3.16 for different IP and OOP conversions.

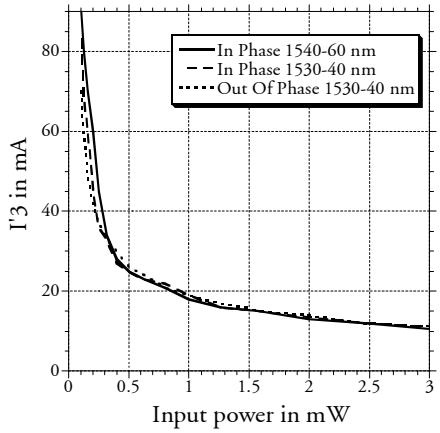


Fig. 3.16: Currents to the input section for "good" operating point at different optical input power levels. Curves are obtained for IP conversion from 1540 to 1560 nm and IP and OOP conversion from 1530 to 1540 nm.

The transfer functions for the different conversions are, as expected, quite similar, but it is noted that the curve is far from linear, which complicates the processing in the control electronics. Furthermore, for "high" input powers the adjustment of the current should be of very high precision in the range of 1 mA/dB.

The sensitivity towards the OSNR level is evaluated by recording the penalty for a BER of 10^{-12} for input OSNRs of 24 and 25 dB/0.1 nm. The result is shown in Fig. 3.17, and it is noted that the probe power to the XGM was fixed to -3 dBm during the measurements.

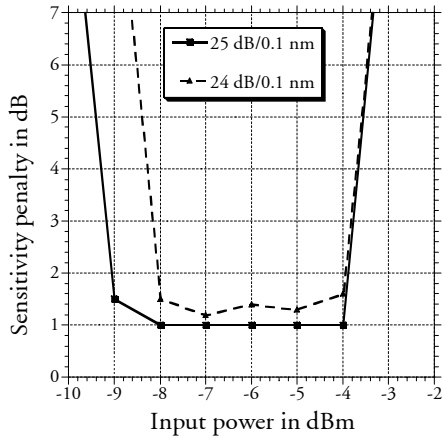


Fig. 3.17: Sensitivity penalty for different OSNR levels when controlling the bias current to the input SOA of the MZI1 stage.

A minimum *input power dynamic range* (IPDR) of 4 dB is deducted from the figure when the OSNR level amounts to 24 dB/0.1 nm. The IPDR is increased to 7 dB, when the OSNR increases to 27 dB/0.1 nm (not shown). It is thus questionable, whether the approach will meet the specifications listed in section 3.3.1.

Moreover, a non negligible dependence of the input packet OSNR to the SOA current is observed for OSNR levels within the specifications. This dependence is probably associated with the steep slope of the transfer function for low input powers in Fig. 3.16. Slight addition of noise requires significant adjustment of the bias current to ensure error free operation. This is a major drawback as the control electronics is not able to detect the OSNR level.

Hence, the requirement for high precision electronics able to cope with non-linear transfer curves, questionable IPDR and injection current dependence towards OSNR levels all cancels the interest of controlling the preamplifier in the input section of the MZI.

3.3.4. Controlling XGM probe power

An alternative to controlling the preamplifier in the MZI input section is to ensure power-equalised packets before the MZI structure, i.e., the packet should be power-equalised after passing the XGM.

This is achieved by controlling the current to the CW laser used as the XGM probe source in order to control the gain. This approach is depicted

in Fig. 3.18 showing the input interface to the regenerator. The non equalised packets experience different gain, and the result is power-equalised packets after OOP conversion and filtering.

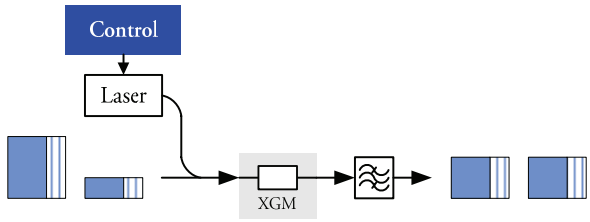
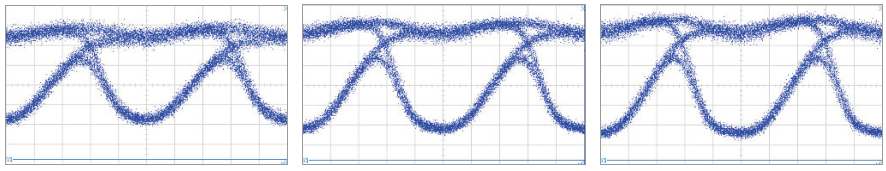


Fig. 3.18: Controlling the optical probe power to the SOA working as XGM

The probe laser can be either distributed feedback (DFB) laser or an *integrated laser module* (ILM) type. To obtain equalised packets at the output of the XGM, the probe power should be linear proportional to the optical input power as described by Equation (3.1).

$$P_{probe} = kP_{in} \tag{3.1}$$

With the same measuring conditions as for the previous approach in section 3.3.3, the packet input power over probe power ratio is manually fixed by controlling the CW laser. The evolution in the eye-diagram, while keeping the ratio fixed to a value of 2 is illustrated in Fig. 3.19. It is seen that the eye-diagrams remain clear and open for an IPDR of more than 10 dB.



P_{in} = -10 dBm P_{in} = -2 dBm P_{in} = +2 dBm

Fig. 3.19: Eye-diagrams obtained at the XGM output for different input powers and probe powers corresponding to Equation(3.1). Time scale: 20 ps/div and vertically: 0.5 mW/div optical power..

The sensitivity penalty is depicted in Fig. 3.20 and an IPDR of more than 7 dB is deducted even for an OSNR of only 23 dB/0.1 nm. The IPDR is increased to more than 12 and 13 dB for OSNR levels of 24 and 25 dB/0.1 nm, respectively. On the other hand, the penalty of 1.5 dB is slightly higher than for the other approach.

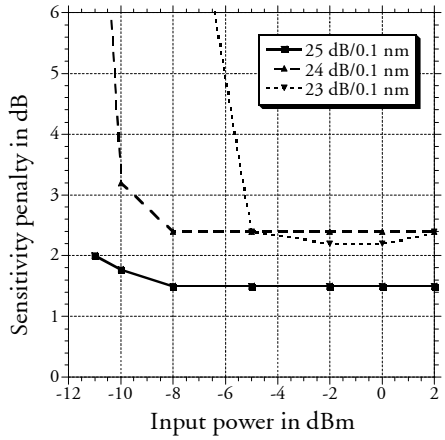


Fig. 3.20: Input power dynamic ranges when controlling the probe power to the XGM.

The large IPDR and the linear transfer characteristic between input power and probe power make the approach highly relevant. In addition, no OSNR dependent effects on the probe current have been observed.

The drawback of this approach is related to the architecture of the packet router. To reduce the component count in the OPR, the XGM can be integrated with the second SOA in the wavelength selector block (see Fig. 3.9). This requires that the control electronics measures the power before the SOA, which consequently requires that an electronic front-end is added for each wavelength.

However, for the DAVID 3R application, the advantages of controlling the XGM by far outnumber the drawbacks, and it is thus obvious that this approach is the best and only viable choice.

3.4. Power Equalisation Control Electronics

In the previous section, it was concluded that the control of the current to the CW probe laser before the XGM was a suitable choice for equalising the power. These evaluations were based on manual control based on the average optical power. In this section the concept is assessed within a packet environment with packet-to-packet power variations.

This requires the presence of control electronics to measure the incoming optical power and determine the current to the probe laser for each time slot.

3.4.1. Requirements to electronic control system

A number of requirements apply to the electronics. It should be quite fast, i.e., it should be able to adjust the current to the laser within the guard band of 50 ns. Then, it should keep the current fixed throughout the duration of the packet length. The power level of the packet should be determined by the power in the header as long sequences of zeroes in the payload should not result in adjustment of the probe laser. The combination of these two requirements cancels the interest for fast SOA based solutions as those suggested in [45], [46], [47] and even [49]. As the electronics are operating as part of a synchronised packet switch, knowledge of the packet rhythm is available, i.e., information of when a new slot arrives. This is very beneficial for resetting before tracking the power of a new packet.

Although the CW probe laser power should be linear to the optical input power, the transfer function between the current and the optical output of the laser is not necessarily linear. For an ILM a linear range of around 10 dB was observed, while the transfer function for a chosen DFB laser was only approximately stepwise linear. Consequently, the control electronics has to compensate for these non-linearities.

Lastly, the receiver sensitivity of the control electronics should be as high as possible, as to reduce the required optical power to the SOA gate.

Hence, the basic requirements to the control electronics can be listed:

- Fast adaptation within the duration of the guard band (50 ns).
- Output should be stable throughout the packet time slot.
- Allowing for generic choice of lasers to control (ILM/DFB).
- High receiver sensitivity.

In the next section, the design of the control electronics taking into account these requirements will be described.

3.4.2. Design of control electronics

To match the requirements in the preceding section, it was chosen to design the control electronics with a digital core in order to add some intelligence in the system. This requires that the detected signal is converted to the digital domain as depicted in the block diagram in Fig. 3.21.

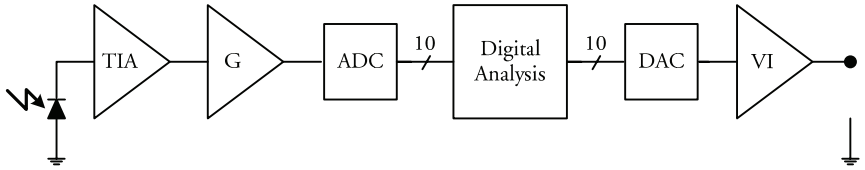


Fig. 3.21: Block diagram of control electronics

A fraction of the optical signal is received by the photo-diode and amplified through the low noise *transimpedance amplifier* (TIA) and the limiting amplifier with gain G. The photo-diode is a cost-efficient pin-diode designed for 622 Mbit/s data reception, and the TIA allows for a maximum bit rate of 2.5 Gbit/s. Additional amplification and filtering are obtained when passing the succeeding low pass amplifier with an upper frequency of 40 MHz.

The result is a fast varying envelope of the packet, which is converted to the digital block with a sample rate of 100 Msps. Here the input is used to settle the output as is explained in the following section. It is noted that the envelope of the packet used for the analysis is independent, whether the packet bit rate is 10 Gbit/s or, e.g., 40 Gbit/s.

The remaining part of the control electronics converts the digital output to a voltage in the *digital to analog* (DAC) circuit, and further to a current suitable for controlling the laser.

In addition to the flexibility offered by the digital design, this approach is considered cost-efficient in terms of component cost and reduced power consumption. All the components are fully commercially available at a cost at the time of writing negligible to the cost of the optical components. The two most expensive parts are the FPGA and the printed circuit board. These parts, however, can be fully integrated with a larger control system of the packet switch. The delay in the electronics is compensated for the optical packet by a delay fibre.

In Fig. 3.22 a photo of the control electronics board is shown, where the analog front-end accounts for the photo-diode, the TIA and the low pass amplifier. The control electronics were developed and implemented at COM•DTU.

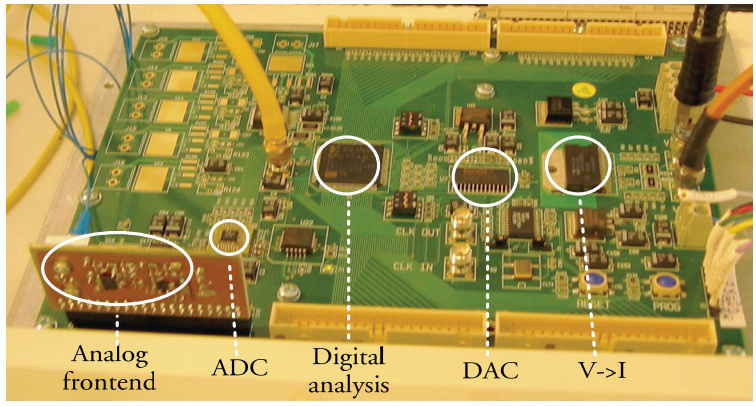


Fig. 3.22: Photo of control electronic.

3.4.3. Digital implementation

The core of the electronic control system is built digitally to cope with the fast adaptation and to obtain independency of the pulse pattern. This also allows for easy reprogramming, as the digital core is implemented in an FPGA through *VHSIC Hardware Description Language* (VHDL) coding.

The digital block is shown in Fig. 3.23, and it comprises three sub-blocks. All the functions can be enabled or disabled by dip-switches on the printed circuit board.

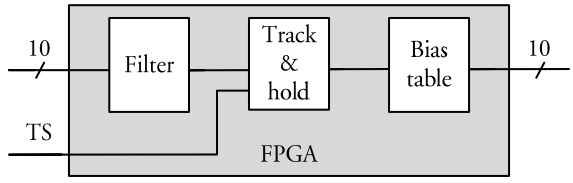


Fig. 3.23: The digital block comprise a digital filter, a track and hold circuit and a bias table to convert between transfer functions.

The digital filter is a first order infinite response (IIR) low pass filter designed to ease the subsequent tracking and holding. Enabling the filter reduces disturbance from the pulse pattern on the envelope, but a delay is introduced, which should be compensated by additional fibre delay line for the optical signal while processing. For the DAVID packet format with a relatively long preamble, it was chosen to enable the filter for the measurements and demonstrations.

The Track & Hold (T&H) block uses two signals for determining the power of the pulse; the filtered power envelope and the timeslot (TS) or packet rhythm signal generated by the packet switch. The TS indicates the beginning of a new time slot; this initiates the tracking process which determines the power level. Several schemes for this have been assessed, and one of the candidates is an evaluation of a specified number of samples after the TS signal; then the largest value is chosen as the packet power. After determining the power of the packet this value is stored and used as output until next time slot. In this way the output is adjusted within a single clock cycle of 10 ns (clock frequency of 100 MHz), which is well within the duration of the guard band of 50 ns.

In the case, where a linear relation does not exist between the current and the output power of the probe element, the current from the control system should be pre-compensated. Hence, the bias-table block contains a lookup table, which serves two purposes. First, the data from the ADC through the filter and the T&H typically has a certain slope and offset. Then the bias table converts to the offset and slope required for the laser to obtain a fixed ratio between the optical packet input power and the probe power. Secondly, if the transfer function for the laser is non-linear, the bias table is programmed to compensate for this by implementing, e.g., a step wise linear transfer function so the resulting output of the laser is linear to the optical input power.

3.5. Experimental verification

As discussed in the previous sections, the approach controlling the current to the probe laser was chosen for implementation. This, in combination with digital and flexible control electronics, is considered as the optimal choice for the 3R regenerator within the framework of DAVID.

In this section, the experimental setup for characterising the approach is described, and it is verified that the input power dynamic range satisfies the requirements listed in the specifications. Finally, noise build-up within the guard band between consecutive packets has been observed. The reason for this is analysed and its impact on the regenerator performance is discussed.

3.5.1. Experimental setup

To verify the performance of the 3R regenerator with power equalisation, it is required to test the device in a packet oriented environment. In addition to the usual test pattern generator, device and error counter, it is required to add equipment for generating packet to packet power variations. In

Fig. 3.24, the experimental setup for emulating a packet environment is depicted. The test pattern generator is programmed to generate packets in the DAVID packet format. Then every second packet is removed through the following SOA device with a gating signal synchronised with the pattern generator. After gating, the signal is split into two branches; in one branch a delay of one packet is introduced, and variable attenuation is added to the other. This approach gives the most realistic packets with respect to SNR. The recombined signal is amplified through an EDFA before entering the regenerator and power equaliser, in which the electronics control an ILM probe laser for the XGM. Finally, a bit error counter measures the quality of the regenerated signal.

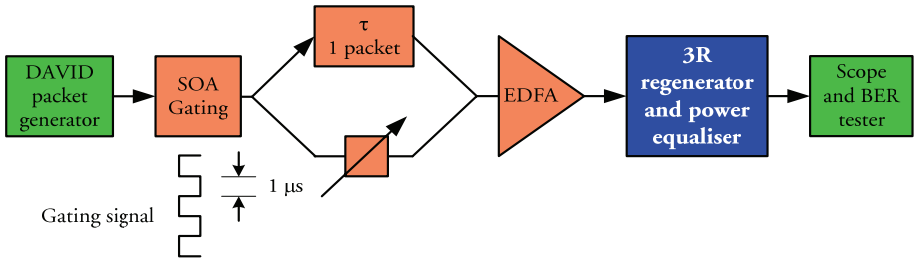


Fig. 3.24: Experimental setup. A continuous stream of DAVID packets is generated, and every second packet is removed in a gating SOA. The signal is divided, delayed, attenuated and combined before EDFA amplification before the 3R with power equaliser.

With this experimental setup, it is possible to generate a difference in the packet power of more than 20 dB, which is far beyond the 5 dB dynamic range in the specifications. The packet power at the output of the gating SOA and the attenuation determines the OSNR of the packets, as noise is added in the EDFA.

3.5.2. Dynamic range

The input power dynamic range (IPDR) was assessed in the environment with packet-to-packet power variations as described above. The eye-diagrams, for different imbalances in power between consecutive packets with the control electronics in operation, are shown in Fig. 3.25.

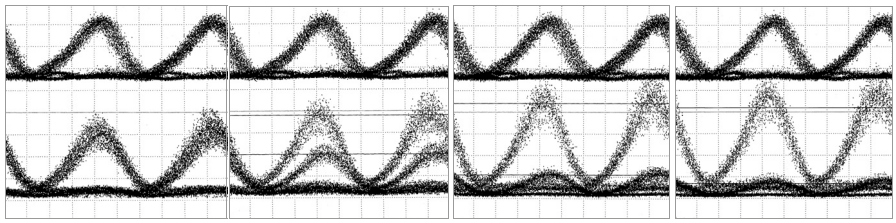


Fig. 3.25. Eye-diagrams for the input (lower) and the output (upper) for different variations in the packet-to-packet power level. Output eye-diagrams obtained after first MZI stage. The time scale is 20 psec/div. and the vertical scales are equal for all measurements.

The lower diagrams show the packets at the input. For equalised packets, the eye-diagrams of the two consecutive packets completely overlap. On the contrary, for a large dynamic range of 9 dB, the spaces of high power packets contain more power than the marks of a low power packet.

The results in the upper eye-diagrams are recorded at the output of the first SOA-MZI stage (MZI1) and indicate that the output of the regenerator is completely independent of the IPDR for variations up to 9 dB. These results are further validated through BER measurements as depicted in Fig. 3.26 for back to back, equalised packets and power variations of 3, 6 and 9 dB. The solid lines are inserted for equalised packets and for power variations of 9 dB.

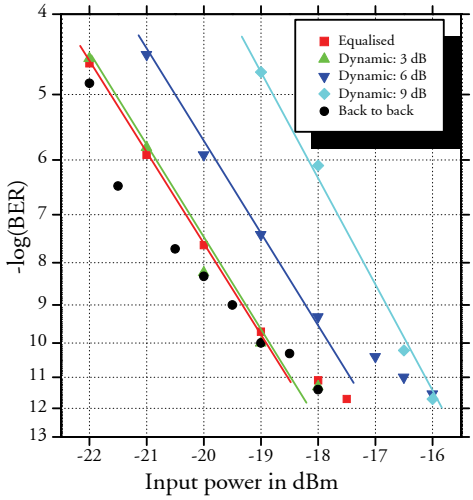


Fig. 3.26: BER measurements with back-to-back and different packet-to-packet power variations from 0 to 9 dB.

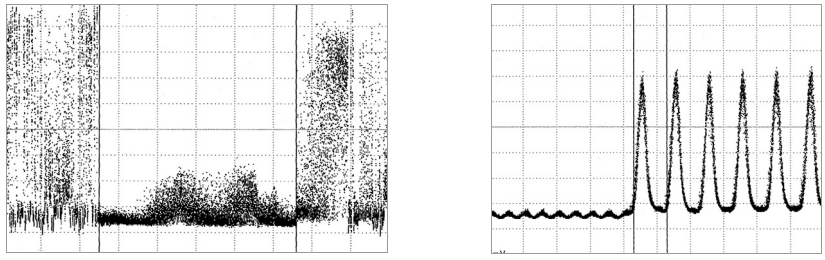
A power penalty of approximately 2.5 dB is observed between equalised packets and a 9 dB dynamic range. Furthermore, no BER floor is observed in any of the measurements, while it is believed that the reduction in the slope for some of the measurements is due to the receiver and the clock recovery as discussed in the next section.

A few bits in each packet were masked due to the receiver before the BER equipment. Under the assumption that the first bits, the preamble, of the packets are unaffected by regeneration, these results documents that an IPDR of 9 dB is achieved. This assumption is verified in the next section, which discusses the behaviour of the guard band during the regeneration.

3.5.3. Guard band distortion

In the BER measurements in the previous section, the first 20-30 bits of each packet were masked as a consequence of the specific clock recovery in the burst mode receiver in front of the BER counter. This is based on the assumption that the integrity of the first bits is preserved; otherwise the results are void.

Therefore, in Fig. 3.27 the guard band after regeneration is shown. The patterns are recorded after the first MZI stage for a packet-to-packet power variation of approximately 4 dB. An examination of the guard band in Fig. 3.27(a) reveals the appearance of distortions from clock pulse residue; however, this has no influence on the first bits of the preamble as documented by Fig. 3.27(b), which shows that the integrity of the preamble is preserved. Similar behaviour is observed for other power differences.



(a) (b)

Fig. 3.27. The guard-band (a) and the preamble of the equalised packet (b) with a dynamic range of approximately 4 dB. The GB in (a) has a duration of 51 ns, while the preamble in (b) has the time-scale of 200 ps. pr. division.

The distortions, however, might influence the cascadability of the regenerators if the power build-up increases. A solution to suppress this effect is to

use a “switched clock” for the retiming stage, i.e., within the guard band the probe power to the first SOA-MZI stage is switched off. This is achieved by logic AND between the clock signal to the SOA-MZI and a guard band suppressing signal synchronised by the switch.

Although the switched clock was never physically implemented, it is believed that this would ensure cascadability of the regenerator

3.6. Summary

In an all-optical switch in a packet network, the packets may arrive with different power level on a packet-to-packet basis. This complicates optical 3R regeneration, where published approaches primarily have been used for regeneration of continuous data streams.

In this chapter the focus has been on the design, development and implementation of control electronics to be integrated with optical 3R equipment for equalising the power level variations

First, however, the advantages and drawback of electrical and optical regeneration was briefly discussed. For electrical regeneration, the main problems are speed limitations and increased power consumption when the bit-rates exceed 40 Gbit/s. Although these limitations do not disqualify electrical regeneration, the focus was directed to optical regeneration, where bit-rates up to 160 Gbit/s have been shown. At the same time, SOA based MZI architectures are potentially very attractive components for optical 3R regeneration, as it is possible to integrate these with an optical switch as the one in DAVID.

The 3R structure adopted comprises an SOA operating in the gain saturation regime and one or two MZIs; the SOA stabilises polarisation differences and reduces the power variations slightly, and the MZIs increase the extinction ratio and reduce the jitter through a clocked probe signal. These two functions are essential for performing regeneration.

Two approaches for controlling the power equalisation were assessed. The first was controlling the gain of the preamplifier in the input section of the MZI. While this scheme looks attractive, it suffers from a non-linear transfer function, high sensitivity towards noise in the control current, and dependence towards the OSNR level. The other approach is based on controlling the gain in the SOA operating as XGM. This is achieved by increasing the probe power for high optical input powers, thus reducing the gain and vice versa. The latter scheme is the most optimal choice of the two because of complete linear transfer function between probe power and input power,

tolerance towards noise in laser current and independence of the OSNR level.

It is chosen to build the core of the control electronics digitally, as this increase the flexibility towards non-linearities and allows for fast adaptation to new packets without compromising long sequences of spaces. The digital part comprises three sub-stages; a digital filter, a track and hold circuit and a lookup table to compensate for any non-linearities in the output.

The performance of the control electronics integrated with the optical 3R regenerator was measured in a packet environment optionally with more than 20 dB power difference between consecutive packets. It was shown that it was possible to equalise packet-to-packet power variations up to 9 dB, which is far beyond the initial requirement of 5 dB. Lastly, the quality of the guard band between packets was evaluated and a build-up of noise was observed. This is due to clock residue, which can be avoided by using a switched clock as probe signal to the MZI stage, as it is documented that the preambles of regenerated packets are completely unaffected by the noise.

The developed, implemented and experimentally verified scheme is a strong candidate for power equalisation and regeneration in future optical packet switched networks. 3R regeneration with power equalisation is a prerequisite for building all-optical networks as the one proposed within the DAVID consortium. This chapter has documented the viability of these functions through experiments at 10 Gbit/s. However, the control electronics uses only the power envelope and is thus completely independent on the bit-rate, why the power equalisation is also feasible at bit-rates of 40 Gbit/s and above.

4. Label processing without header modification.

The processing of a packet in the MPLS framework usually requires that the MPLS label is modified because the labels only have local significance. In optical packet switched networks it is thus required to read and erase the packet header, modify the header and synchronise the new header on the payload of the optical packet. This is far from a mature and available technology.

Several lab implementations of optical label swapping have been demonstrated. Some of them use the label and the payload on the same wavelength separated in time by a guard band. This was the main assumed approach in the DAVID project, and an overview of basic techniques for dealing with optical headers is provided in [50]. These techniques usually require a guard band between the header and the payload. This is, however, overcome in [51] where the optical modification is done directly on the header through a Mach Zender device without erasing the header. This approach still assumes bit-level synchronisation, which is challenging. Other implementations with orthogonal labelling are studied in several projects [52] and the technique is further presented in numerous papers, e.g., [53].

Hence, conventional label assignment schemes for MPLS require header modification in the core nodes of the network because the labels have only local significance. While this is of minor concern for electrical switch nodes, it is a major challenge for optical nodes. An alternative to MPLS in the optical layer is standard IP, where header modification is not essential[1]. However, traffic engineering with IP requires signaling protocols and pre established connections, why the scalability is questionable.

In this chapter a novel scheme for processing in the core nodes without header modification is proposed. The scheme, the Key Identification Scheme, is presented in section 4.1 including a description of the algorithm that the scheme is based on. The scheme imposes some requirements to the network, which could affect the scalability. This is considered by simulation and analysis in section 4.2. The integration of the scheme with MPLS networks and suggestions for an FPGA implementation is treated in section 4.3 and 4.4, respectively. Finally, concluding remarks are given in section 4.5.

---

[1] The *Time To Live* (TTL) and the CRC fields are usually adjusted for each node or each domain (ATM)

In the chapter, which is primarily based on [Publ. 1] and [Publ. 5], the general terms *edge* and *core* nodes are used to identify the routers in the network. These do correspond to the E-LSR and the T-LSRs in the MPLS terminology, respectively.

4.1. Key Identification Scheme

The use of the *Key Identification Scheme* (KIS) is motivated by the limited functionality in the optical layer. Instead of a solution in the physical layer, the KIS focuses on an addressing scheme based on a well known algorithm to determine the output port in the core nodes. Basically, a label is created and this, and only this, label is used for addressing the output ports in the core nodes.

First, in section 4.1.1 the concept of the KIS is presented, and then the *Chinese Remainder Theorem* (CRT) is derived. While the CRT in section 4.1.2 is presented in mathematical terms, section 4.1.3 addresses the use of CRT in a network application. Finally, an example of the label computation is given in section 4.1.4.

4.1.1. Concept of addressing scheme

The basic idea in KIS is sketched in Fig. 4.1, where a packet is transported through a network.

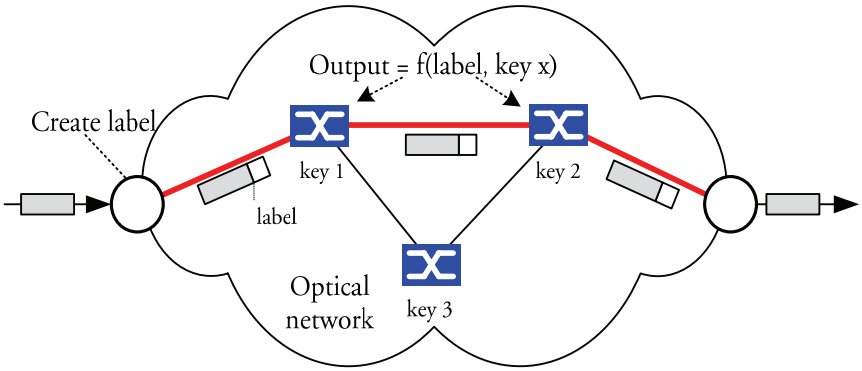


Fig. 4.1: Basic concept of KIS. The label is added at the edge node, and the output port is determined by a function on the label and a local ID.

As in MPLS, a label is added at the E-LSR, and this label is used to designate a unique path through the network. At each of the core nodes (T-LSRs) the forwarding decision, i.e., finding the output port, is given by a mathemati-

cal function based on the label and a node-unique key. In the figure, the output of node 1 is determined by the result of the function f(label, key1). Hence, the packet with the unmodified header is forwarded to node 2, where the same operation is performed and the output of node 2 is given by f(label,key2).

In the following section, an algorithm for implementing KIS is introduced, i.e, the algorithm is an example of implementing the function f(label,key).

4.1.2. Chinese Remainder Theorem

The Chinese Remainder Theorem (CRT) was published in the first century by the Chinese mathematician Sun Tsu Suan-Ching. The basic idea is to represent an integer by the remainders left after division by a set of integers. The famous problem was to find an integer which left the remainders 2, 3 and 2 when divided by 3, 5 and 7, respectively [54]. CRT was the algorithm that was used to find the lowest positive integer $a = 23$ as an answer to the problem.

This has been used for converting between decimal numbers and the *Residue Number System* (RNS), where big integers are represented by arrays of residues. This has shown advantageous in high speed *Very Large Scale Integration* (VLSI) design as it allows for parallel processing of multiplications and additions [55].

The same properties are utilised for implementing the Key Identification Scheme. A path is described in RNS by two arrays; one array is defined by unique IDs (keys) for the nodes en route, and the other array is formed by the corresponding desired output ports for each of the nodes. With these two arrays, the CRT is used to compute a unique label with the properties as sketched in Fig. 4.1.

The array with the keys and the array with the output ports are denoted $\overline{n}$ and $\overline{a}$, respectively. Hence, for a path of k nodes, the two arrays are given:

$$\overline{n} \leftrightarrow (n_1, n_2, ..., n_k) \text{ and } \overline{a} \leftrightarrow (a_1, a_2, ..., a_k) \tag{4.1}$$

The following calculations based on [51] use the arrays in (4.1) to compute the label, which is denoted as the scalar a.

First, a new scalar n is defined as the product of all k elements in the array $\overline{n}$.

$$n = n_1 \cdot n_2 \cdot ... \cdot n_k \tag{4.2}$$

Secondly, a new array $\overline{m}$ is created where the elements satisfy following equation:

$$m_i = \frac{n}{n_i} \quad \text{for } i \le k \tag{4.3}$$

This ensures that each element in array $\overline{m}$ satisfies $m_i \bmod n_j = 0$ for $j \neq i$, i.e., m_i is the product of all elements in array $\overline{n}$ except n_i.

A third array $\overline{c}$ is created based on array $\overline{m}$ and array $\overline{n}$:

$$c_i = m_i (m_i^{-1} \bmod n_i) \quad \text{for } i \le k \tag{4.4}$$

The term m_i^{-1} is called the *multiplicative inverse* of $m_i \bmod n_i$ defined by the relation $(m_i^{-1} \cdot m_i) \bmod n_i = 1$. This is valid if m_i and n_i are relative primes, i.e., they should have no common divisors larger than one ($\gcd(m_i, n_i) = 1$). Because m_i is the product of all *n*-elements except n_i, all elements in array $\overline{n}$ should be relative primes to satisfy the requirement. The multiplicative inverses are easily calculated using the GCD recursion theorem as described in Euclid's *Elements* (~300 B.C.).

Finally, the scalar *a* is calculated based on array $\overline{a}$ and array $\overline{c}$:

$$a = (a_1 c_1 + a_2 c_2 + \ldots + a_k c_k) \bmod n \tag{4.5}$$

Given the scalar *a* and the array $\overline{n}$ it is straightforward to restore the array $\overline{a}$ using following expression:

$$a_i = a \bmod n_i \quad \text{for } i \le k \tag{4.6}$$

It is thus seen that the above algorithm based on CRT can be used to construct a scalar *a* from two arrays $\overline{a}$ and $\overline{n}$, where all elements in the latter are pairwise relative primes. Given the scalar *a* and the array $\overline{n}$ each element of $\overline{a}$ is uniquely restorable.

In KIS the scalar *a* is transported as part of the label, and each node en route uses equation (4.6) as the function to extract or decode the forwarding information.

4.1.3. Forwarding based on CRT

In MPLS networks, the required forwarding information at each node should at least include addressing the output port. Optionally, information about class of service etc. could be included either as path-unique properties or as node-unique properties. In the latter case, this information should be included for each node.

This is illustrated in Fig. 4.2, where the label comprises node unique and path unique information. The node-unique information is to be extracted for each node, while the path-unique information is shared between all nodes on the requested path.

Hence, each node decodes the node-unique part of the label to retrieve the *Forwarding Information Field* (FIF), which comprise information about output addressing and other relevant node-unique information. The figure also indicates the flexibility of the scheme as the FIF for different nodes do not need to be exactly the same size.

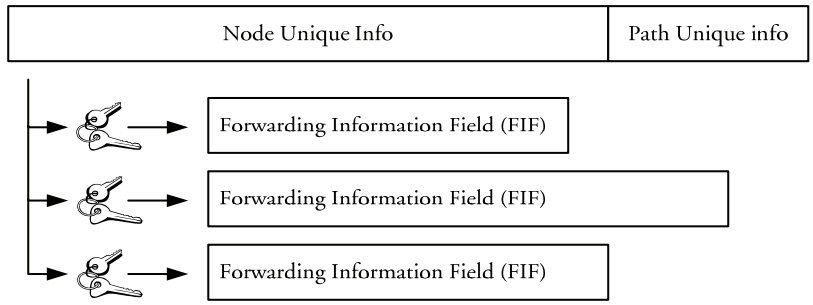


Fig. 4.2: The keys are used to extract the node specific FIF from the label.

In the following, for simplicity, the label only includes the encoded node-unique info, why equation (4.6) can be rewritten to:

$$FIF_{node} = \text{Label mod } key_{node} \tag{4.7}$$

As it was discussed in section 4.1.2, the requirements to the keys are that they should be unique, and that they should be pairwise relative primes, i.e., their greatest common divisor should equal one. Furthermore, a key for a node should be larger than the largest size of the FIF that it is used to decode.

Considering an MPLS based network the KIS based on CRT is implemented in following five steps.

1) When the network is configured, each node is assigned a key satisfying the above requirements. The keys are distributed with proprietary signalling or as piggybacking to existing signalling.
2) Information of the topology of the network and the assigned keys are provided to the edge devices. This is obtained with conventional routing protocol like, e.g., OSPF.
3) For all announced paths through the network, a label is computed using CRT. It is noted that this is only done in the edge devices.

4) When a packet enters the network, its destination is read, and the corresponding label is added to the packet before launching it to the network.
5) Each core device or T-LSR should use equation (4.7) to identify the output port address and other relevant information.

It is noted that the existence of any label distribution protocol like LDP, CR-LDP or RSVP is not required. This again reduces the requirements to the control part of the core nodes, as they do not need to support label distribution components. In section 4.3 the similarities and differences between the KIS and MPLS are discussed more thoroughly.

4.1.4. Label computation example

As an example, consider the transmission of packets through a network supporting the KIS.

It is assumed that keys are assigned to the nodes in the network, and that the route through the network is advertised to the edge node using conventional routing protocols. Hence, the requested route comprises three core nodes with the keys 25, 14 and 37, which are pairwise relative primes although only 37 is an absolute primes. As illustrated in Fig. 4.3, the output ports for the three nodes are 4, 2 and 3, respectively.

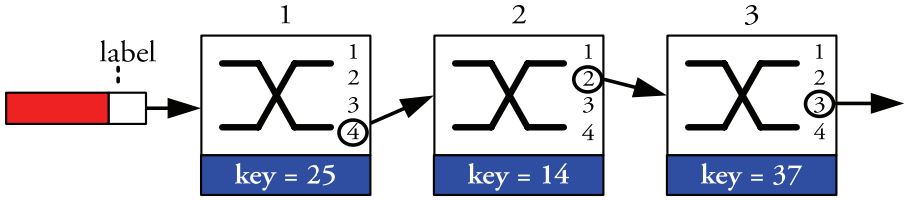


Fig. 4.3: A path through the MPLS network comprises three core nodes. The output ports are 4, 2 and 3 while the keys for the nodes are 25, 14 and 37.

The third step is the computation of the label using the algorithm presented in section 4.1.2.

The two arrays $\overline{a}$ and $\overline{n}$ are populated with the output addresses and the keys for the nodes:

$$\begin{aligned} \overline{a} &= \text{outputs addresses} = (4,2,3) \\ \overline{n} &= \text{keys} = (25,14,37) \end{aligned} \tag{4.8}$$

With these arrays it is possible to calculate $n = 25 \cdot 14 \cdot 37 = 12950$, the array $\overline{m}$ and the multiplicative inverses:

$$\overline{m} = (518, 925, 350) \tag{4.9}$$

$$\begin{aligned} 518^{-1} \bmod 25 &= 7 \\ 925^{-1} \bmod 14 &= 1 \\ 350^{-1} \bmod 37 &= 24 \end{aligned} \tag{4.10}$$

The results in (4.10) are easily verified:

$$\begin{aligned} (7 \cdot 518) \bmod 25 &= 3625 \bmod 25 = 1 \\ (1 \cdot 925) \bmod 14 &= 925 \bmod 14 = 1 \\ (24 \cdot 350) \bmod 37 &= 8400 \bmod 37 = 1 \end{aligned} \tag{4.11}$$

Based on equation (4.4) and the multiplicative inverses in (4.10), the array $\overline{c}$ is computed:

$$\begin{aligned} \overline{c} &= (518 \cdot 7, 925 \cdot 1, 350 \cdot 24) \bmod 12950 \\ &= (3626, 925, 8400) \end{aligned} \tag{4.12}$$

Finally, the scalar a is calculated as described by equation (4.5):

$$\begin{aligned} a &= (4 \cdot 3626 + 2 \cdot 925 + 3 \cdot 8400) \bmod 12950 \\ &= 41554 \bmod 12950 = 2704 \end{aligned} \tag{4.13}$$

According to step 4 in section 4.1.3, a label of 2704 is added to the packet, after which the packet is launched to the network. The fifth step refers to the decoding in the core nodes as described by equation (4.7), where the FIF equals the output port address.

$$\begin{aligned} &\text{Node 1: } 2704 \bmod 25 = 4 \\ &\text{Node 2: } 2704 \bmod 14 = 2 \\ &\text{Node 3: } 2704 \bmod 37 = 3 \end{aligned} \tag{4.14}$$

Hence, the label 2704 can be used to forward the packet through the network path shown in Fig. 4.3.

4.2. Scalability of KIS

In the previous sections, it was assumed that the node specific keys should be pairwise relative primes. It is, however, an open question how this will impact the performance of KIS, when the network size increases.

This question is addressed in this section, where simulations are performed on randomly connected networks. Initially, it is chosen to restrict the evaluations to network sizes less than 50 core nodes. This choice is based on

an investigation of a large number of backbone networks by different service providers [56], indicating that 10-30 core nodes are sufficient. A similar number is used in the European ACTS project OPEN [57], which considered an all-optical network based on optical cross connects.

After discussing the simulation results, an analysis of networks up to 200 nodes is given.

4.2.1. Simulations scenarios

To address the scalability of KIS using CRT, a large number of randomly connected networks is constructed by computer simulation. Generally the networks are characterised by following properties:

- All interconnections are full duplex, i.e., if a connection exists from node *x* to node *y*, then a similar connection exists from node *y* to node *x*.
- A number between 1 and 6 is randomly chosen for each node. This number defines the number of connections that is created to other core nodes in the network.
- The network integrity is ensured as possible "single islands" are connected with full duplex connections, i.e., in case a randomly connected network turns out to comprise two or more separated networks, these are connected using extra duplex connections.

The model for generating the random network topologies was implemented in C++. As an example, in Fig. 4.4 two randomly connected networks generated by the model are given.

Both networks comprise ten core nodes, but the topologies have turned out very differently, as the network in (A) is considerably more densely connected than the one in (B). This illustrates the variety in the simulated networks, and it is noted, that the length of a path equals the number of hops, i.e., all the links have unified weight.

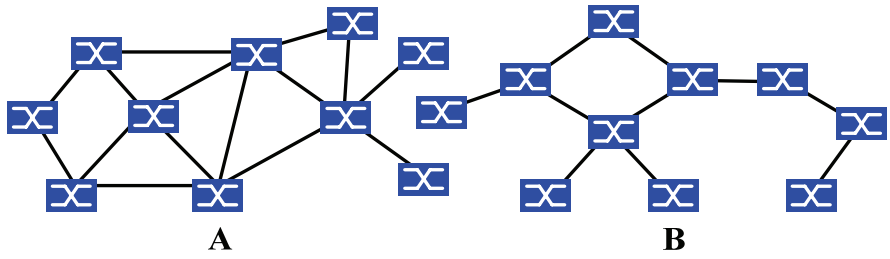


Fig. 4.4: Examples of randomly connected networks generated by the model.

Two different scenarios with different distributions of node sizes[1] in the networks are considered as sketched in Fig. 4.5. The uniform distribution refers to a network where values of 8 (50%) or 16 (50%) are required in the FIF, and the non-uniform distribution refers to networks with a wide range of node sizes ranging from 4 to 256. Note, as discussed in section 4.1.3 that the size of the FIF does not need to match the number of output ports.

These two node distribution scenarios are chosen to give an indication of the robustness of the scheme against variations in the node and FIF sizes.

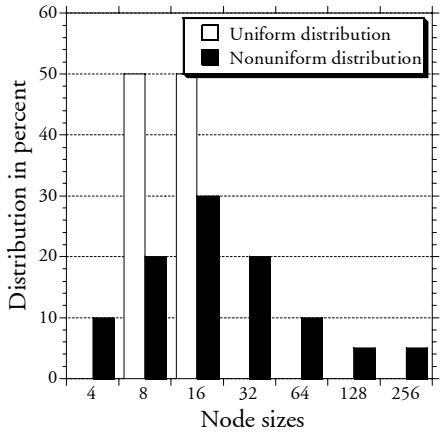


Fig. 4.5: Scenarios with uniform and nonuniform node size distribution.

4.2.2. Simulation results

For each generated network, valid keys were applied to the nodes with two restrictions. Firstly, the key ID should be larger than the size of the FIF it is used to decode as described in equation (4.7). Secondly, the key should be unique and should have no common divisors with any previously assigned keys in the network.

The shortest path between any node pairs was discovered using Dijkstra's single-source shortest path algorithm [51]. Then the label size for each path was computed using worst case equation (4.2). For each network, two values were stored, the average and the maximum label size. The first is an average of all label sizes, and the latter is the maximum of all label sizes in the network. 20.000 network topologies were generated and averages of the two

[1] The "node sizes" are understood as an "overlay" to the number of adjacencies in the model, i.e., the node sizes (port numbers) are distributed to the nodes independent on the actual topology of the generated network.

values were obtained. These are in the following denoted *average* and *maximum*. The large number of iterations ensured convergence of the results.

The result for different network sizes are shown in Fig. 4.6(a) for the distribution of FIF sizes as given in Fig. 4.5. It is seen that the results from the two scenarios are almost identical. This is especially clear for increased network sizes, and it indicates that the addressing scheme is robust against a non-uniform node size distribution. Furthermore, to support a network of 50 nodes a label size of 6-7 bytes is required depending on whether the average or the maximum lengths of the paths in the networks are chosen.

Fig. 4.6(b) shows the results for prioritised key distribution, i.e., the most used nodes in the network are assigned the lowest possible keys. This is implemented by initially resolving all shortest path through the network to determine which core nodes are used most intensively. The required label size is reduced with half a byte for uniform distribution and a few bits for non-uniform distribution compared to the results for unprioritised key distribution. hile prioritisation of the keys reduces the label size requirements slightly, the robustness towards non-uniform node size distribution is compromised. Furthermore, minor updates in the network topology, node failures and traffic engineering might require transmission through lower prioritised nodes with higher keys. This will inherently require larger labels sizes, why this simple prioritisation does not seem to be advantageous.

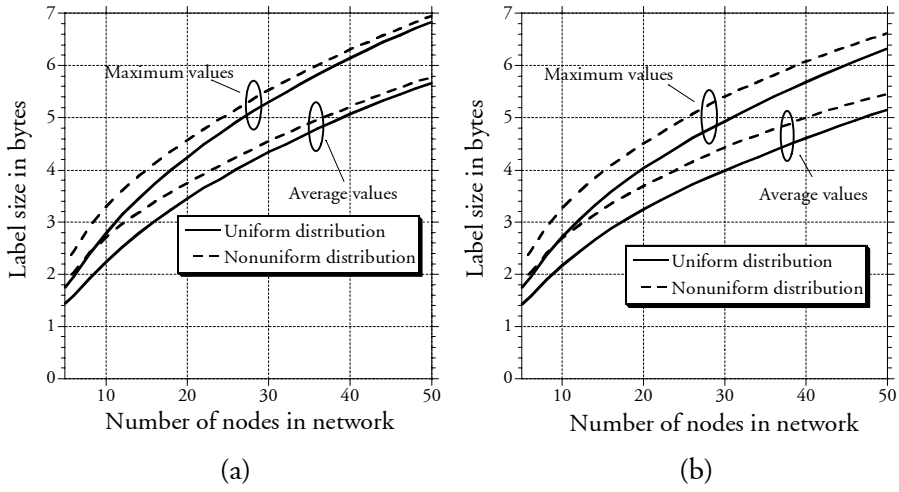


Fig. 4.6: Required label size as function of network size. In (a) the assignment of keys to the nodes are done randomly, while prioritisation is used in (b).

4.2.3. Analysis for very large network

Even though the main focus is on networks with less than 50 core nodes, it is relevant to analyse which factors that influence the scalability for very large network, e.g., up to 200 core nodes.

The increase in the label size is dependent on two factors. First, larger randomly connected networks tend to include longer routes between a pair of nodes, i.e., more information has to be encoded with the same label. Secondly, the available keys are limited to the relative primes, which increase non-linearly with the network size, making the average valid key larger.

The first factor is analysed by measuring the path length for all possible node-to-node pairs in the network. The result is shown in Fig. 4.7, and with the given assumptions for the network topology, it is seen that the distribution of route lengths moves slightly towards longer paths for an increased network size. Hence, the average path length is approximately 4, 5 and 6 hops for a network size of 50, 100 and 200 nodes, respectively.

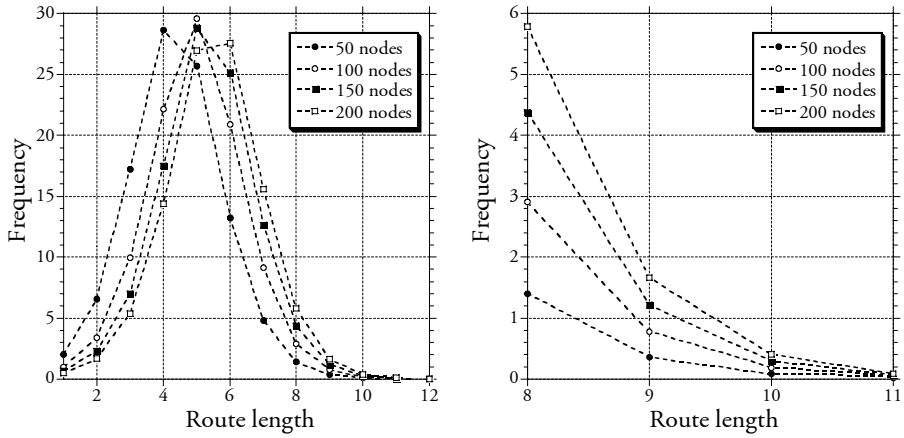


Fig. 4.7: Overview and zoomed for long routes of distribution of path length in the networks.

Fig. 4.7 is furthermore useful to indicate the impact of reducing the maximum allowed path length. Consider a maximum path length of 7 hops. For a network size of 100 nodes, this requirement will affect about 2.9+0.8+0.3+0.1 = 4.1% of the paths, which is computed by a summation of the frequencies for routes lengths from 8 to infinite.

The second factor is evaluated by computing average key sizes for different networks sizes. For a uniform network, the size of each key depends on the

available primes. Therefore, as the size of the available primes increases nonlinearly, the average key size will increase accordingly. As an example, the average key size in a network of ten nodes with FIF sizes of 8 is calculated:

$$\text{avg. key size} = \frac{1}{10}(9+10+11+13+17+19+23+29+31+37)=19.9 \qquad (4.15)$$

Thus, the estimated label size is computed as the path-length times the average key size.

$$\text{avg. label size} = \text{ path length } \cdot \text{ avg. key size} \qquad (4.16)$$

This estimation is depicted in Fig. 4.8, where the required label size is sketched as function of the network size for different path lengths.

It is, e.g., seen that a label supporting path lengths of 6 nodes in a network comprising 100 nodes will require a label size of around 6 bytes. The fractions shown with the circles are derived from Fig. 4.7(b). They indicate the number of path for a given network size that has a hop count of 9, 10 or 11 nodes, e.g., for a network size of 100 nodes < 1.0% of the paths are longer than eight nodes.

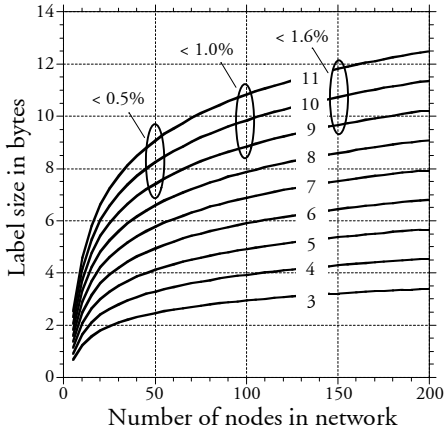


Fig. 4.8: Required label depending on the network size and the path length. The circles denote the number of paths of length 9, 10 or 11.

The analysis in Fig. 4.8 is compared to the simulated results in Fig. 4.6 for a network size of 50 nodes. According to the analysis only less than 0.5% of the paths are longer than eight hops, why a restriction to this is acceptable, and Fig. 4.8 shows that this requires a label size of 6-7 bytes. For the same network size the simulations in Fig. 4.6, for the *maximum* values, show that

7 bytes are required. This small difference is explained by the contribution from the 0.5% paths longer than eight hops, why the results in the simulation correspond closely to the analysis.

Hence, it is indicated through Fig. 4.8 that in case the path length is restricted to 8 hops, a network of up to 200 nodes can be supported by the KIS with 9 bytes in the label.

The small fraction of paths that do not satisfy the length requirements are addressed in section 4.2.4.

It is relevant to compare the topologies generated by the model with real topologies as observed in the Internet. It turns out that the topologies in the Internet tends to be less random than the ones in the model [58][59]. This reduces the longest path, why the simulations in this chapter can be considered as a conservative estimate.

4.2.4. Scaling impact on network performance

If the label becomes too large to fit in the header, the path through the network has to be divided into two or more shorter paths. This is illustrated in Fig. 4.9, where a path from A to B is routed through node 1, 2, 3, 4, 5 and 6.

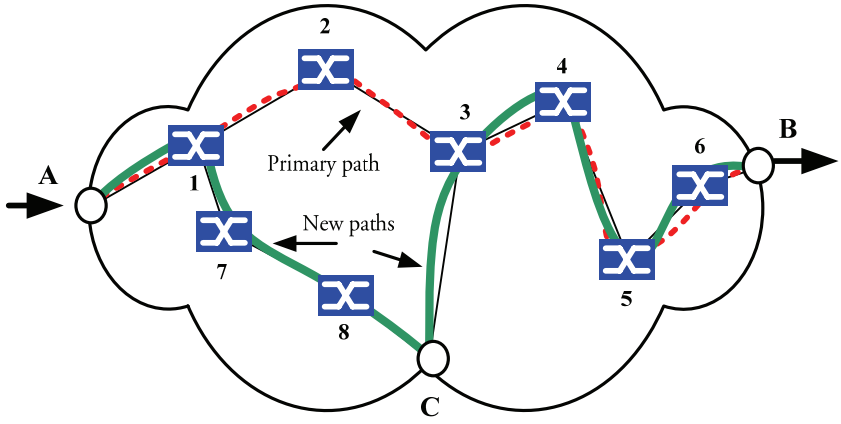


Fig. 4.9: The path AB is subdivided in AC and CB.

Assuming that the label for the path becomes too large, the path AB is subdivided in AC and CB as depicted with the solid line.

The procedure for subdividing a path into sub-paths is as follows. When the routing entity in A resolves a route which requires a too large label, it re-

solves a new route to an edge node C closer to the destination than A. Then the label for AC is pushed to the packet and the labelled packet is transmitted through nodes 1, 7 and 8. At C, the inner label or network layer header is used to determine the route to B through nodes 3, 4, 5 and 6. The new routing decision is thus based on the network or IP header.

As the example illustrates, division of a path generates more overall traffic and processing in the edge nodes. However, if only a negligible fraction of the paths experience subdivision, the amount of extra processing and traffic is accordingly negligible. Alternatively, the network can be divided into different hierarchies to reduce the number of nodes at each level as it was illustrated in Fig. 2.8 in section 2.2.3. The impact of division into sub domain has been considered in [60], where numerical simulation results indicate a reduction in the label length of approximately a factor of 2 compared to KIS.

4.3. Integration of KIS and MPLS

This section addresses the integration of the KIS with MPLS networks, as was discussed in chapter 2.2. Part of the chapter is based on [Publ. 6], which considers a network comprising several MPLS domains.

In Fig. 4.10 a GMPLS network is shown, which includes smaller islands of GMPLS, MPLS or KIS domains.

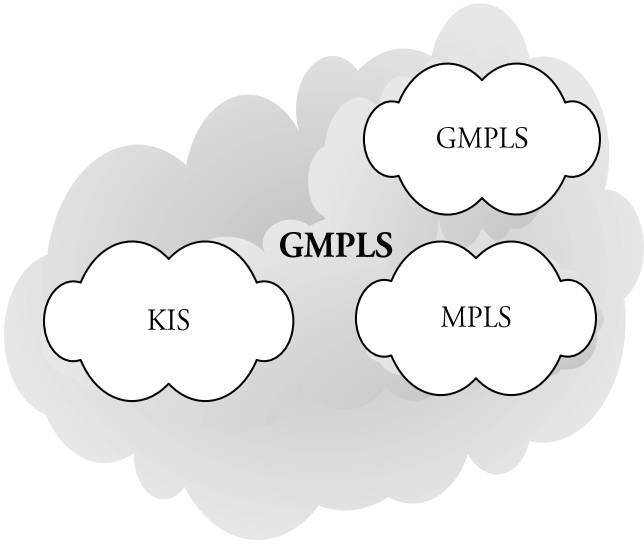


Fig. 4.10: Integration of KIS with in a (G)MPLS network

The KIS is compliant with the MPLS framework with a separation of the forwarding and routing components. Thus, all routing functionality is pushed to the edge of the network and the core nodes are only responsible for forwarding the labelled packet using only local information. The KIS label is pushed to the existing label stack at the edge node regardless whether this label stack is empty, e.g., an unlabelled IP packet, or whether the stack contains several labels from an MPLS networking hierarchy.

In Table 4.1 the main procedures and responsibilities for KIS and standard MPLS are listed. The table is separated both for the control and the data plane and for edge and core nodes.

Table 4.1: Comparison between standard MPLS and KIS

		MPLS	**KIS**
Control Plane	**Edge**	• Label distribution protocol for path establishment or manual configuration. • Resource reservation with RSVP-TE, CR-LDP etc. • Explicit and independent routing supported	• Labels computed using CRT • No resource reservation. • Only explicit routing • All control in edge
	Core	• Label distribution protocol • Maintaining forwarding table	• Unique key
Data plane	**Edge**	• MPLS label added based on LSP/FEC	• KIS label added for each FEC
	Core	• Lookup operation • Label swapping	• Modulo expression • No header modification

Seen from the data plane in the edge node, the only difference between KIS and MPLS is the format and the size of the label. Both labels associate to a path for a given FEC. The main difference from the data plane is in the core nodes. Whereas in standard MPLS, the label is used for looking up the new label and the output port, in KIS the output port is computed using modulo operation.

For the control plane in the edge node, the label distribution protocol in the standard MPLS is basically substituted with the CRT label computation of KIS. Furthermore, no resource reservation in KIS is included and only explicit routing is supported. The control operations in the core nodes are

heavily reduced for the KIS compared to MPLS, as the control entity solely comprises the unique key as introduced in this chapter.

Seen from the higher layers in an MPLS hierarchy or from the network layer, the differences in the control of KIS or standard MPLS are negligible. Therefore, the KIS can operate as a layer in an MPLS hierarchy.

4.4. Implementation of the KIS

In this section, the implementation of the KIS in the core nodes is addressed. The main function of the control system is to compute the output port based on the label and the key, as described in the previous sections. This is based on modulo calculation, why efficient implementation of this is considered.

First, architectures for calculating the remainder after integers division are presented, and implementation details are considered. Then, the scalabilities of the suggested solutions are evaluated.

4.4.1. Modulo operation

The implementation of modulo operation, which is required in the core nodes, is basically the well known problem of conversion from binary to residue numbers [61][62]. The problem is formulated in the following equation, where x is the remainder. Note that $|A|_M$ denotes $A \bmod M$.

$$x = A \bmod M = |A|_M = A - M\lfloor A/M \rfloor \tag{4.17}$$

As seen, the remainder or residue is calculated using a division, a rounding function, a multiplication and a subtraction. This is, however, very time consuming in an FPGA due to the complexity of the division operation. Therefore, the following considerations aim on exchanging the multiplication and division operations with a simple addition modulo M [63][64].

Consider the usual binary representation of the label scalar A, where $a_i \in \{0,1\}$ for $i < n$ and $n = \lceil \log_2(A+1) \rceil$:

$$A = (a_{n-1}, a_{n-2}, \ldots, a_1, a_0) \tag{4.18}$$

This is identical to the summation:

$$A = \sum_{i=0}^{n-1} 2^i a_i \tag{4.19}$$

This expression for A is then used with equation (4.17):

$$x = |A|_M = \left| \sum_{i=0}^{n-1} 2^i a_i \right|_M = \left| \sum_{i=0}^{n-1} \left|2^i\right|_M a_i \right|_M \tag{4.20}$$

The last term is valid as $(a+b) \bmod n = (a \bmod n + b \bmod n) \bmod n$ [51]. As all values $\left|2^i\right|_M$ are known, x is straightforwardly calculated by addition modulo M as discussed in the following section.

4.4.2. Implementation of modulo adders

As shown in the previous section, finding the remainder after integers division is decomposed to an n-operand addition modulo M, where n is the number of bits in the scalar to be divided and M is the divisor.

A two-input modulo M element can be implemented as a usual adder circuit and a succeeding comparator to test whether the sum is greater than M, in which case M is subtracted from the result. For n operands, the addition is implemented as a tree with depth $\lceil \log_2 n \rceil$ of 2-input modulo M adders. The n "known" values are pre-computed externally and stored in registers with a size equal to the size of M.

As an example, a label size of 8 bytes (A) is used with a 1-byte key (M). This requires 64 table lookups for the $\left|2^i\right|_M$ (for $i \in \{0,1,...,n-1\}$) values. The result is a 64 operand 8 bit modulo M adder, which is implemented in 6 stages.

In Fig. 4.11 two different implementations of modulo adders are shown. Each bit in the label (64 bits) is used to indicate whether the corresponding lookup table value should be included in the addition modulo M.

In (a) the addition in implemented as a full 64 operand modulo M adder, whereas in (b), the addition uses two steps. On the contrary, in (b) the addition modulo M is replaced by a standard 64 operand adder tree, which produces a sum of maximum 14 bits, and this 14 bits word is then used in the next step to look up 14 values for the 14 operand modulo M adder. The two stage adder in (b) is proposed to reduce the area and power consumption of the circuit, as will be evaluated in the following section.

An all-optical implementation of the Key Identification Scheme has been suggested in [65]. This implementation is based on Terahertz Optical Asymmetric Demultiplexers (TOAD), which according to simulations performs modulo operation.

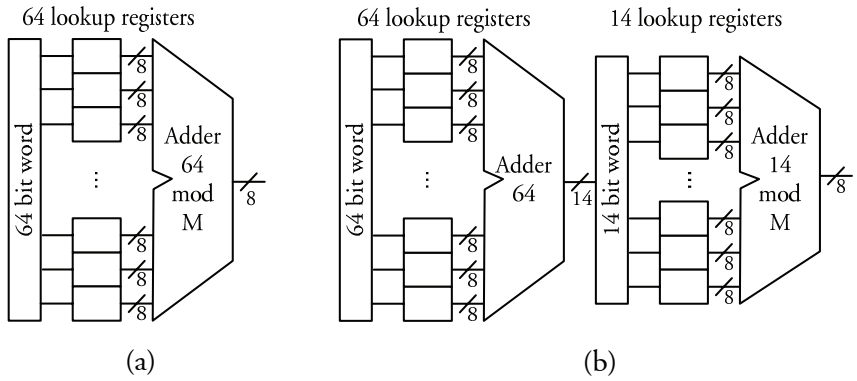


Fig. 4.11: Diagram of modulo adder based on lookup tables. In (a) the calculator comprises a 64 operand modulo adder, whereas in (b) a 16 operand modulo adder is used together with standard 64 operand adder.

4.4.3. Scalability of implementation

The suggested implementations for calculating the remainder should be evaluated with respect to the delay and chip-area requirements.

Hence, the two proposed implementations in the previous subsection are evaluated by measuring the delay and space requirements for supporting 16, 32 and 64 bit label sizes, respectively. This is achieved by implementing the circuits in VHDL for the Xilinx VirtexE FPGA technology. The results of the synthesis are shown in Table 4.2, and it is noted that listed frequencies and delays do not include wire delays from the place and routing process. Despite this, the results are useful to evaluate scalability and for comparison between the two implementations.

Table 4.2: Delay and area requirements for modulo implementations. The area values are normalised with 16 bit mod 8 bit single lookup as reference

	Single lookup table		**Two lookup tables**	
Size	**Cycles/Frequency**	**Area**	**Cycles/Frequency**	**Area**
16 bit mod 8 bit	5 / 173 MHz	*100*	6 / 161 MHz	140
32 bit mod 8 bit	6 / 168 MHz	202	6 / 158 MHz	201
64 bit mod 8 bit	7 / 168 MHz	404	7 / 158 MHz	325

The single lookup table version refers to the circuit in Fig. 4.11(a), while the two lookup table versions refer to the one in Fig. 4.11(b). The area consumption for the single step 16 bit mod 8 bit device is used for reference.

For small label sizes, the single lookup solution is superior regarding both delay and space requirements, while for a label size of 64 bits the two stage solution has slightly reduced area consumption.

Most important, however, is that the total delay is far less than the packet duration in DAVID of 1 μs [22], and the area consumption for the largest implementation is in the range of 10% of the size of a standard FPGA.

A key size of one byte is usually sufficient to support networks with up to 50 nodes, as it is possible to assign more than 50 pairwise relative primes smaller than 256 within the network. This might, however, be compromised if the network only comprise very large nodes. Despite this, these implementations clearly indicate that the KIS is realisable for extracting the remainder in the core nodes within the duration of a packet.

4.5. Summary

Header processing in all-optical networks is complicated requiring high precision synchronisation of the new header to the packet. It is thus attractive to avoid header processing as to avoid the bit-synchronisation. Furthermore, the scalability of globally significant source-destination is rather inferior.

Therefore the Key Identification Scheme (KIS) as an approach to avoid header modification in optical networks was introduced and evaluated. The basic idea is to include all the routing information within the label; for each node, the forwarding information is identified by a *key* and the label itself. Furthermore, the Chinese Remainder Theorem (CRT) was derived, and it was shown that the CRT was a suitable algorithm for implementing KIS. Through CRT, a label for a path is created, and this label is used for forwarding the packet through the network. The only operation required for the core nodes is a mathematical modulo operation on the label and the identification key.

The CRT algorithm does, however, impose some requirements to the available identification keys, i.e., all the nodes in the optical network domain should be assigned a unique key, and this key should have no common divisors with any other key in the network. Therefore, the scalability of the scheme was evaluated by simulations and analysis. The simulations were based on random network topologies for evaluating the size of the label.

These simulations all indicated that a label size of approximately 6 bytes was sufficient to support network sizes of up to 50 optical nodes.

Although it is not believed that larger optical networks are relevant in a near future, an analysis was made to identify which factors that influence the scalability. First, the route length tends to grow for larger networks and secondly the available key size increase non-linearly with the network size. The analysis covered network sizes up to 200, and if the maximum route length was set to, e.g., 8 nodes, then a network size of 200 is easily supported. Any path length not satisfying the requirement is subdivided in two or more separate paths adding only a negligible overall traffic and processing.

The integration of KIS with standard MPLS networks was discussed for the data and the control plane. For the edge node, the main difference is that the usual label distribution protocol is replaced by the CRT label encoding scheme

Finally, implementations for the processing in the core nodes were discussed. Using a ROM based implementation with a small number of precomputed values, even an eight byte label was processed well within the duration of the packet. Two different implementations were compared regarding delay and area requirements, and generally a ROM based solution with modulo M adder was preferred.

It is thus shown that the Key Identification Scheme with CRT can be implemented in optical network nodes, and that it can coexist with MPLS networks. Furthermore, the scalability is acceptable, and the header modification else required is avoided by the scheme. This is compromised by a label size that is two to three times the size of a standard MPLS label. Although the scheme applies to both electrical and optical networks, the main focus is on optical packet switched networks, as header modifications in electrical nodes are straightforward.

The avoidance of header modification is a major step towards realisation of optical networks, as this has been one of the major obstacles.

5. Fibre Based Access Networks

The paradigm shift from the information society to the entertainment society is heavily reflected in the usage patterns for the end-users network connections. And vice versa, the availability of broadband connections prepares the soil for new demanding services.

Applications like Video on Demand (VoD), a huge number of high quality TV channels over the Internet, video-phones etc. have strict requirements to the delay, which is hard to satisfy in combination with high bandwidth demands [66]. Other drivers for the increasing the capacity of the access network are central content storage and backup facilities replacing local backup approaches as CD ROMs, DVDs etc. Once the access mile bottleneck is removed it does not matter where the content or backup facility is located whether it is on the local hard drive or on a remote content server.

Fibre access is considered as one of the most promising physical media to support future access network requirements. When and how FTTH will be deployed is, however, dependent on the cost-efficiency of the fibre technologies in combination with the willingness of the user to pay for the offered services.

The objective of this chapter is to evaluate the optical components, which constitute an FTTH system. Especially, it will be discussed whether directly modulated lasers are suitable for fibre access at high bit rates. The evaluations are based on results obtained from a case study.

Section 5.1 details the classic problems of the “last mile” including the market pressure in this segment. The cost-efficiency of the system design and components are discussed in section 5.2 and in section 5.3 a four channel FTTH system at 10 Gbit/s is implemented with focus on the choice of lasers, receivers, channel distribution and electronic front-end. The verification is addressed in section 5.4 followed by a discussion on the component choices for future FTTH systems in section 5.5. Finally, a summary of the work is provided in section 5.6.

5.1. The last mile

The transmission span from the end user to the telephone exchange or any other provider is widely recognised as the famous ”last mile”, and this has

almost always been considered as *the* bottleneck between the user and the network. In this last mile, two communication technologies are displacing others; the wireless technologies enables mobility and fibres offer practically unlimited bandwidth. When this is said it is expected that the recent developments in the xDSL-technologies will copper survive for some more years.

The last mile differs from the core network as only few users share the costs. This increases the importance of cost-efficiency in comparison with the core network, where the cost is shared by all the users. This cost-efficiency, in combination with the user willingness to pay for the new services, determines the pace of the FTTH deployment. These topics are briefly addressed in this section. First, the importance of the cost is indicated by studying the willingness to pay for high bit-rate services. Then, a short view on the evolution on the access technologies is sketched and the differences between passive optical networking approaches are outlined.

5.1.1. Market pressure on access technologies

The implementation of high capacity access technologies is depending on a realistic business plan, i.e., if the customers are not willing to pay sufficiently for new services to cover the equipment investment, no new deployment will happen.

In a study presented in [67], the user willingness to pay for different services are discussed, and a similar approach will be used in the following. Usually, different communication and entertainment services have been linked very closely to a specific physical media, each with its own cost structure. However, with the availability of broadband connections the high bandwidth one way services like television merges with the low bandwidth interactive services like, e.g., phone conversations. In Table 5.1, different services are listed and it is roughly indicated how much an average user is willing to pay for the different services based on the price paid for similar service today. There are factors of more than 10000 between how much a user is willing to pay pr. bit for TV to the home compared to phone conversations. The table provides an indication of the importance of reducing the costs for the last mile access. Without an extremely cost-efficient infrastructure, no deployment of fibre bases access solutions should be expected unless heavily subsidised, e.g., by the government [68].

As stated in [69] the optoelectronic part of the optical access infrastructure is significant enough that the total cost of fibre access deployment will depend on this assuming the fibre *has* been installed. Hence, the main limit for business cases, and thus real deployment, is the cost of the lasers, the multiplexers, the connectors and the receivers. Especially for the compo-

nents located at the customer premises as only one customer should pay for the equipment.

Table 5.1: How much to pay for a bit?[1]

Service	Calculation	Price pr. bit
Telephony	National call: €0.03 pr. minute	9000 p€/bit
Web surfing	Basic 512/128 ADSL: €40 pr. month. 3 hours pr. day of intense surfing	1500 p€/bit
Video on Demand	Rental price in video store: €5 pr. DVD. (2 hours of 5 Mbit/s)	140 p€/bit
Viewed TV	2 TVs 7 hours a day. €30 pr. month (DVD quality)	4 p€/bit
Available TV	20 channels	0.4 p€/bit

5.1.2. Fibre to the premises

The access technologies within the recent years have evolved from using *Plain Old Telephone Service* (POTS)[2] connections with bit rates up to 56Kbit/s, through ISDN terminals to mainly ADSL broadband connections from 128 Kbit/s up to usually 4 Mbit/s[3]. New versions of ADSL, ADSL2+ [70], allow downstream bitrates up to 24 Mbit/s on the copper wires. While the ADSL (and its extensions) technology provides medium bit rate access in distances up to 5-6 km., the VDSL technology provides bit rates up to 200 Mbit/s for very short ranges of a few hundred meters [71][72].

The introduction of fibre access technologies complements the improvements in the DSL technologies as shown in Fig. 5.1, which illustrates three versions of FTTx architectures.

In the FTTCab scheme the *passive optical network* (PON) spans from the *Optical Line Terminal* (OLT) at the *Central Office* (CO) location to an *Optical Network Unit* (ONU) located in a cabinet in close vicinity of the end users. The last few hundred meters to the end user's *Network Terminal* (NT) uses DSL technologies, e.g., VDSL. One ONU would preferably serve several NTs.

---

[1] Source: TDC (Febr. 2006) and a common Danish video rental store

[2] Some confusion exists for this abbreviation. POTS are also used for Plain Ordinary Telephone Service and Plain Ordinary Telephone System.

[3] The ADSL standard supports downstream bit rates up to 8 Mbit/s, however, in Denmark the highest bit rate offered by most providers in 2006 is 4 Mbit/s

In the Fibre to the Building, FTTB, the ONU is moved closer to the end user, and consequently the link between the ONU and the NT can use shorter distance technologies like copper based Ethernet or wireless access.

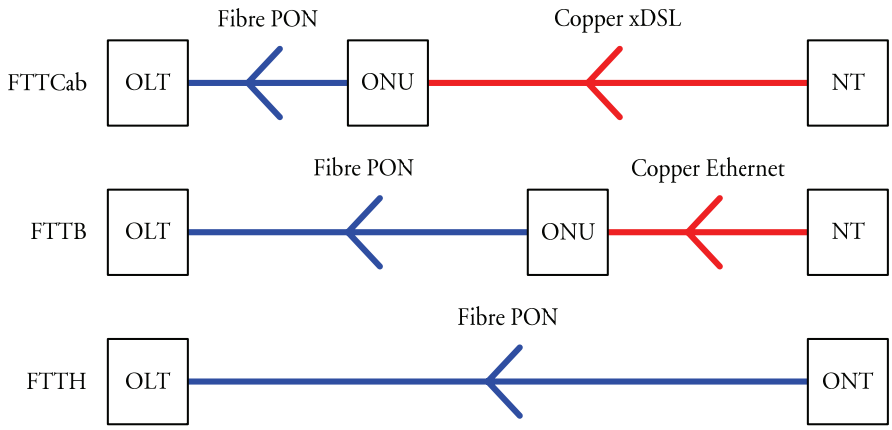


Fig. 5.1: Fibre to the Cabinet, the Building and the Home

Fibre to the Home (FTTH) is defined as the situation, when fibre is deployed on the complete path to the end users premises. Here, the ONU and NT are merged into the *Optical Network Terminal* (ONT). The bandwidth and services offered to the end user now depends only on the capacity of the PON independent of the DSL or Ethernet limitations.

It is expected that the deployment of fibre access in areas with an existing copper infrastructure will evolve from FTTCab through FTTB to FTTH. Such strategy will allow the deployment of the infrastructure even if the take-rate is rather low.

In the remainder of this chapter the terms FTTH will mainly be used although it might just as well denote FTTCab or FTTB.

5.1.3. PON Technologies

In order to share at least some of the costs of deploying FTTH, the *Passive* Optical Network (PON) has been introduced. The PONs use passive optical components as splitters, combiners, couplers and multiplexers for the optical infrastructure in combination with the fibre.

In Fig. 5.2 the general operation of a PON is illustrated. The downstream channel uses 1550 nm wavelength region while the upstream uses the 1310 nm wavelength region.

In the downstream direction from the OLT to the ONUs, the information for the different ONUs is mixed and each ONU receives all the data and passes those destined for its own users or clients.

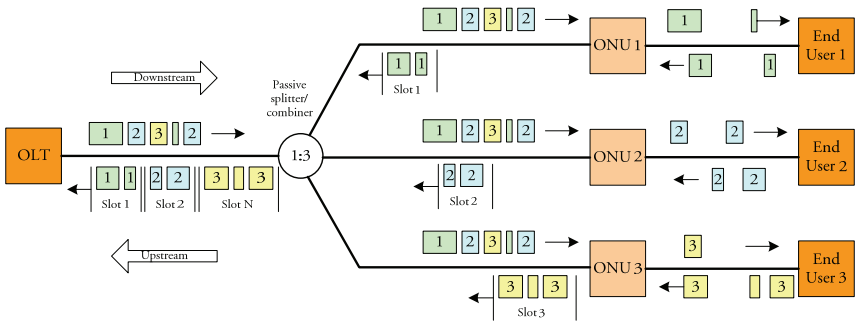


Fig. 5.2: Upstream and Downstream operation for a 1:3 PON. (Inspired by [73])

While the downstream transport is rather straightforward, the upstream traffic needs careful control. No buffers are implemented in the passive 1:3 optical splitter and combiner. Hence, each ONU are granted timeslots that it can use for upstream information transfer. In Fig. 5.2, the end user sends upstream information with no restrictions, but the ONU only forwards the data in granted timeslots. An assumption for the PON scheme is a ranging of the distance between the ONU and the combiner, why either the OLT or the ONU measures the roundtrip time between these two.

The implementation of a PON increases the complexity of the receiver in the OLT due to the variations in phase and amplitude between slots from different ONUs. Hence, the upstream bit rate is usually a factor of four or more lower than the downstream rate.

There exists PONs based on ATM (APON) [74] and Ethernet (EPON) [75], and the reader should consult the references for further information.

5.2. Cost efficient fibre access

In the preceding section, the importance of cost-efficient fibre based access solutions was outlined. The different scenarios for deployment of the fibre to the cabinet, the building or the home were described.

In this section, the focus is on the different components that constitute an FTTH system and how they are interconnected. In this way a point to point scenario is outlined, which is used in the remainder of the chapter to verify different laser types, modulation schemes and wavelength configurations for future high speed FTTH solutions. It is thus important to note that the scope is not to evaluate components and architectures for current

FTTH solutions, rather to provide guidelines and indications for choice of components for future significantly higher capacity systems.

5.2.1. Transmission limitations

Two main factors limit the physical range for FTTH systems; the power budget and the dispersion parameters. This is especially true for FTTH systems with no optical amplification and directly modulated lasers, which is likely to increase the effect of dispersion. These two limiting factors are considered in the following.

5.2.1.1. Power Budget

The power budget indicates whether enough power reaches the receiver to maintain reliable performance of a transmission system. It is furthermore used as an indicator on how little input power that potentially can be accepted for proper system operation.

Including a system margin S, the received power P_{rec} is given by the following expression where C_L is the channel loss accounting for loss in components and fibre and P_{out} is the laser output power of the transmitter

$$P_{rec} = P_{out} - C_L - S \tag{5.1}$$

The channel loss C_L is calculated by the following expression where the α_f is the loss pr. km in the fibre, L is the length of the fibre, α_{con} is the connector loss, α_{splice} is the splice loss and α_{comp} is the insertion loss of the components:

$$C_L = \alpha_f L + \alpha_{con} + \alpha_{splice} + \alpha_{comp} \tag{5.2}$$

The main difference between the point to point system and the PON is the value of the α_{comp} parameter, which for PON includes the "loss" caused by the power splitting. Theoretically, a splitting ratio of 1:64 would increase α_{comp} with 18 dB. The actual loss depends on the quality and type of components.

The expression will be used in section 5.3.2 when different systems are evaluated.

5.2.1.2. Dispersion

Dispersion in a single mode optical fibre originates from chromatic dispersion, i.e., the transmission velocity varies with the different spectral compo-

nents that are included in the pulse. Hence, a narrow spectral width is obviously the best choice to overcome dispersion problems.

The impact of dispersion varies with the wavelength with the zero-dispersion close to 1260 nm and increasing with wavelength. For a 1550 nm laser "far" from the zero-dispersion, the maximum fibre length L is computed using the following expression [1]. B is the bit-rate, D the dispersion parameter and $\Delta\lambda$ the spectral width.

$$L < \frac{1}{B|D|\Delta\lambda} \tag{5.3}$$

Direct modulation of the lasers, as will be discussed in section 5.2.2, increases the spectral width of the emitted light. As an example, a 1550 nm laser modulated at 10 Gbit/s has a spectral width of $\Delta\lambda$ = 0.40 nm [76]. Hence, according to [77] a typical dispersion value for 1550 nm transmission in single-mode fibres is 17 ps/(km nm), which can be inserted in equation (5.3):

$$\begin{aligned} L &< \frac{1}{10.7 \cdot 10^{9} s^{-1} \cdot 17 \cdot 10^{-12} s/(km \cdot nm) \cdot 0.40 nm} \Leftrightarrow \\ L &< 13,7 km \end{aligned} \tag{5.4}$$

Hence, for at laser operating in the 1550 nm region with a spectral with of 0.40 nm, a maximum transmission distance of 13.7 km using standard single mode fibre is indicated.

The problem is overcome by reducing the spectral width of the pulse, by changing the dispersion value of the fibre or by compensating the dispersion through *Dispersion Compensated Fibres* (DCF) The first is hardly possible as the problem is inherent for direct modulation, the second is costly with higher loss factor, and the latter require replacement of the installed fibres with dispersion shifted or dispersion reduced fibres.

As operation in the 13xx nm wavelength is much closer to the zero-dispersion wavelength the problem of dispersion is here of minor concern, however the effect of four-wave-mixing exists for large launch powers.

5.2.2. Components

Reducing the system costs of FTTH systems implies reducing the costs of the components and merging components. A general block diagram of a channel in an FTTH system is shown in Fig. 5.3.

The electrical parts of the transmitter comprise the serialiser and the amplifier. The, e.g., 16 width data stream is serialised into a 1 bit wide stream and the signal is amplifier in the laser driver amplifier before the laser. The fibre network includes the transmission fibre and the optical multiplexing and splitting components depending on the topology of the FTTH network.

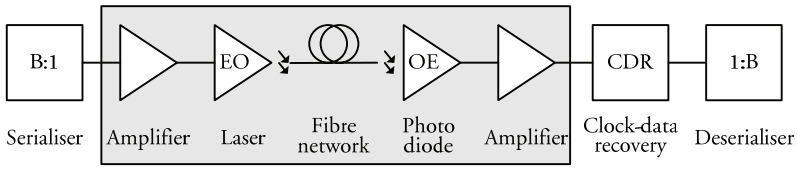


Fig. 5.3: General FTTH transmission system.

The receiver constitutes apart from the photo diode, a series of amplifiers, a clock and data recovery and framing into a B wide data bus.

The following sections focus on the components in the grey box of Fig. 5.3.

5.2.2.1. Lasers

Cost efficient integration and usage of lasers in an FTTH system depends among others on the following factors:

- Design yield
- Integration of lasers and laser drivers
- Wavelengths stability
- Power consumption.
- Optical output power flexibility

The design yield defines the "success rate" in the production of the devices, i.e., the ratio of components from a badge, which operates and satisfies the required specifications. Evidently, relaxing the requirements to the component, i.e., allowing a simpler design, improves the yield.

The integration of several lasers and/or laser drivers is expected to lower the costs and increase the performance compared to separate components. A high yield is here even more important when several components are integrated on the same substrate.

The wavelengths selected for a FTTH system should stay within the pass bands defined in the ITU-T CWDM grid [78], which is 13 nm with a channel spacing of 20 nm.

The power consumption of the lasers (and laser drivers) is directly reflected in the maintenance costs. Apart from an increased power usage, cooling is

required at the components and centrally at the central office. This should be avoided; however, still a sufficient optical output power is required to cope with transmission and splitting losses.

The above constraints influences the choice of the lasers and laser technology, and three important selection parameters are listed below:

- Modulation format
- Wavelength selection
- Cooling

The lasers can be modulated in two ways; direct modulation and external modulation.

In directly modulation the current modulates the intensity of carriers in the laser cavity, which is directly reflected in the number of created photons and the laser power. However, the rapid change in the carrier density modulates the refractive index in the laser, thus modulating the phase and frequency of the transmitted pulses. This creates chirped pulses, which are more likely to experience group velocity dispersion in the transmission fibre as discussed in section 5.2.1.2.

Alternatively, external modulation can be utilised. Here a continuous wave laser is providing a narrow signal, which is modulated by an external modulator using, e.g., interferometric techniques. The result is a spectrally very narrow signal suitable for long haul transmission. The drawback, on the other side, is power-consuming modulation and for EA[1]-modulated signals a quite low optical modulated output power.

The power and complexity properties of direct modulated lasers are inherently superior to the external modulates lasers. However, the influence of chirp induced dispersion should be evaluated.

The price of a 15xx nm DM laser is currently significantly larger than the price of a 13xx nm laser. It is, however, unknown whether this is simply due to the lack of mass-production of 15xx nm lasers for direct modulation, which prevent them from long haul transmission use. Lasers in both wavelength bands should be included in the evaluation.

Generally the threshold current of lasers depends on the temperature increasing exponentially with increased temperature. This, according to [1], normally allows operation at temperatures up to 50 to 70 degrees C, and operation is hardly possible above 100 degrees. However, as it is described in the literature, e.g. [79], a lot of experiments and commercial products are

[1] EA: Electro Absorption modulation.

verifying high temperature operation in the 13xx nm region. Based on the literature, it is slightly indicated that lasers in the 13xx nm region are less sensitive to temperature variations opposed to their 15xx nm region counterparts. The impact of removing active cooling of the lasers should be considered in the case study.

5.2.2.2. Photo diodes

The photo diode converts the optical input power to an electrical current. The PIN diode is a simple semiconductor converting photons to electrons, thus creating a current proportional to the received light. An extension to the PIN diode is the avalanche photo diode (APD), which have a considerably higher sensitivity, but as well a much higher noise figure. In the APDs the recombination into electron-hole pairs introduces further recombination as defined by a multiplication factor *M*. Hence, there is a trade-off between the increased complexity and noise associated with the APD compared to the improved sensitivity.

In order to reduce the overall power consumption and relax the requirements to the laser diodes, a receiver with a high sensitivity and low noise is preferred. While the absolute theoretical minimum power to obtain a Bit Error Rate (BER) of 10^{-9} with a 10 Gbit/s signal is -48,6 dBm [1], Table 5.2 shows the sensitivity figures for commercial state-of-the-art diodes.

Table 5.2: Commercial photo diodes and their sensitivity

Manufacturer	Model and type	Bit rate	Sensitivity
Bookham	PIN – PT10XGC	10 Gbit/s	-20.5 dBm @ 10^{-12}
Bookham	APD – AT10XGC	10 Gbit/s	-28.5 dBm @ 10^{-12}
Bookham	APD – AT3SGCC	2.5 Gbit/s	-35.0 dBm @ 10^{-12}
JDS Uniphase	PIN – ERM568	10 Gbit/s	-19.5 dBm @ 10^{-12}

The inferior performance of the receivers compared to the theoretical limit is caused by imperfect responsitivity, thermal noise, timing jitter and the fact that no input signal shows an infinite extinction ratio.

The extinction ratio is improved by operating the lasers, where the slope of the I-P curve is steepest. Together with high extinction ratio of the laser drivers this increases the power level in marks compared to the power level in the spaces. In a FTTH system where no optical amplification is intro-

duced, the only significant noise contributors are the laser and the receiver. It is thus noted that none of the noise contributions significantly depends on the length of the fibre span, which thus should be limited only by the dispersion figures and the power budget as discussed in section 5.2.1.

In the case study, only a simple PIN diode is used. This is mainly because of the more simple design, which is believed to improve the yield.

5.2.2.3. Frontend electronics

The frontend electronics constitutes the electronics required to convert the digital signal to a laser current for the laser and in the receiver to convert the small current from the photo diode into a digital signal.

The digital logic is usually developed in a CMOS process, which have significant advantages with respect to high yield and low power consumption. The latter is closely connected to the low voltage swing required in CMOS. It is thus desired to integrate the front-end electronics in CMOS technology if possible.

However, as the supply voltages of the modern CMOS processes are reduced to save power in very complex digital designs, this impacts the current that can be launched from an integrated laser driver. Such issues may require higher speed and power consuming technologies like SiGe and GaAs. This makes monolithically integration on a single chip impossible and a multi chip solution would be required.

In the receiver path, the use of low noise electronic equipment and proper shielding of the electronic parts is important as the current level from the photo diode is small. The use of small optical input power levels dictates transimpedance amplifiers (TIA) with very low noise figures. Again the monolithically or at least very dense integration is crucial to avoid noise and RF implications. The conversion in the TIA should be followed be a non-linear amplification through CMOS limiting amplifiers, which delivers an output signal sufficient for the digital logic.

5.2.2.4. Passive optical components

This section addresses general considerations on the choice of passive optical components like (de)multiplexers, couplers, splitters and circulators.

In PON solutions the simple power splitter or 1:N coupler shown in Fig. 5.1 is required to distribute the optical power through the downstream distribution network.

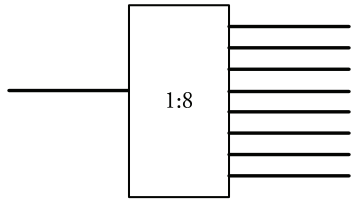


Fig. 5.4: 1:8 power splitter.

As the component within the recent years has been mass produced the price has been reduced accordingly. Furthermore, the passive optical components provide bit rate transparency, and the same component can be used in a 155 Mbit/s and a 10 Gbit/s system.

The inherent drawback of the splitter is the power loss due to the splitting ratio. E.g., theoretically the loss is 9 dB for a 1:8 splitter and commercial products indicates losses of approximately 11 dB [80].

The couplers do not differentiate between the different wavelengths in the fibre, and thus it is not sufficient for a FTTH system with several wavelengths that should be treated separately.

In contrast to the coupler, the *Coarse Wavelength Division Multiplexer* (CWDM) separates and multiplexes different wavelengths as shown in Fig. 5.5. This is required for multi-wavelength systems, where the wavelength domain is used for either increasing the overall capacity or for distributing wavelengths to different users.

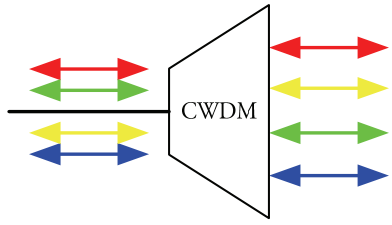


Fig. 5.5: Wavelength multiplexing/demultiplexing with CWDM device

Principally CWDM devices have been in use since the early 1980s [81], however, the term "CWDM" was not used in the industry before 1996. The CWDM was invented as a low cost supplement for DWDM networks in the metro area, and it was the intention to allow for variations in centre frequency etc. Furthermore it covers all the wavelength range from the O-band, through the E, S and C-band and to the L-band. This is very suitable for operation with G.652.C fibre, where the water peak in the E-band is negligible, and the full wavelength range can be used.

The spacing between the channels is 20 nm. This spacing is chosen as a compromise between occupying only a little bandwidth and for allowing uncooled lasers to operate from 0 to 70 degrees.

The CWDM devices are fabricated using a thin-film filter technology, which compared to arrayed waveguides (AWG) require much fewer layers, which significantly improve the yield and reduce the cost. The insertion loss in commercial CWDM devices today is about 0.6 dB per channel [82].

Using the same wavelength for communication in both directions requires an optical circulator, as sketched in Fig. 5.6. Here, only an insertion loss of approximately 1 dB is introduced for light from 1 to 2 and 2 to 3. However, from 1 to 3 and 2 to 1 a suppression loss of 40-50 dB is present [83].

In addition to the passive optical components, the connectors and splices also contribute to the channel loss. Depending on the quality of the connector, this introduces 1-2 dB loss and splices approximately 0.2 dB. These figures are approximate values and the real loss depends on the connector type, the environment and the skill of the technician on site.

It is, however, obvious that splices should be preferred compared to connectors whenever possible due to the significantly lower loss value.

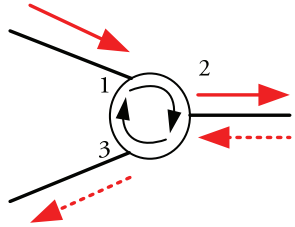


Fig. 5.6: Optical circulator. Light from input 1 is directed to output 2, and from input 2 to output 3.

5.3. FTTH solution at 10 Gbit/s

The theoretical considerations were evaluated through a case study. The overall objective of the study is to determine the viability of directly modulated lasers for future high speed FTTH deployments. It is initially decided that only point-to-point solutions should be considered, which simplifies the system design considerably as no burst mode transceivers at 10 Gbit/s are required.

The implementation goal is to develop a system with two bi-directional channels of each 10 Gbit/s. In this way one channel is reserved for HDTV

(and SDTV) applications with possibility of interactivity while reserving another 10 Gbit/s channel for 10 Gigabit Ethernet transmissions. This is sketched in Fig. 5.7, where the 10 G home box is installed at the end user, while the 10 G central box is installed in a local or regional central. The two boxes are connected through one single fibre connection of approximately 10 km. This allows easy and flexible deployment of the system, as only single fibre infrastructure is required.

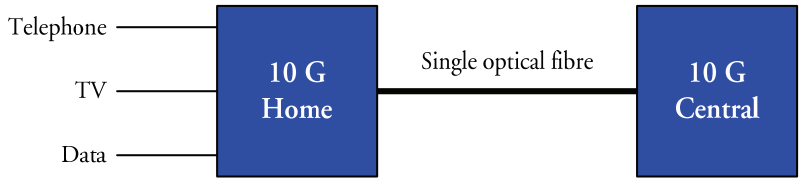


Fig. 5.7: Overview of 10 Gbit/s FTTH system requirements

At the home user, the 10 G box provides interface to the required services like TV, data and *plain old telephone system* (POTS). At the central, on the other hand, the traffic is aggregated onto a MAN or WAN network, which is not specified further here. Note, that even the system is named a 10 G system it is in fact a two times 10 Gbit/s full duplex transmission system.

The only way to considerably reduce the costs of the system is by ensuring a high degree of integration of the components on few chips. This is illustrated in Fig. 5.8, where the two grey areas denote the goal for the integration with all the electronic equipment in few CMOS chips and all the optical equipment on an InP substrate. The functions and devices of each of the chips are sketched.

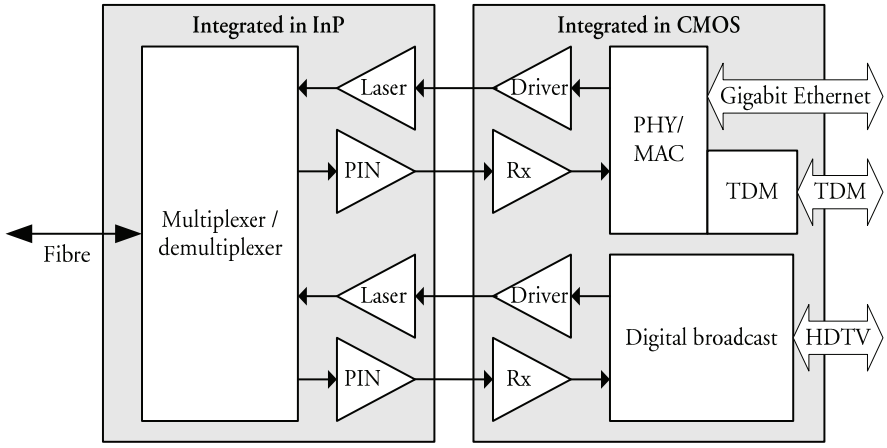


Fig. 5.8: Integration of the digital system and the analog frontend in few chips.

The Rx blocks comprise TIA, limiting amplifier (LA), clock and data recovery and framing. The TIA converts the current from the photo diode to a voltage, which is amplified in the non-linear limiting amplifier. The clock information is extracted from the signal and used to recover and reframe the signal into a 16 bit parallel signal. The driver on the other hand serialises the signal and boosts to drive the laser

In the following sub sections the requirements to the system from different aspects are highlighted.

5.3.1. Case study requirements

As described in the previous section, the basic requirements to the developed case study can be outlined:

- Single fibre system due to connector cost and flexibility
- In total four unidirectional channels are required. Two in each direction.
- The optical frontend should allow a bit rate of 10.7 Gbit/s. This is 10 Gbit/s with overhead for Reed Solomon(255,239) FEC encoding [84].
- Long term integration and cost reductions should be considered.
- Simple design, e.g., directly modulated lasers.
- Compliance with national and international safety standards.
- Low power consumption

In the following, the requirements from a system and an integration perspective will be detailed.

5.3.1.1. System requirements

The maximum allowed BER has been chosen to 10^{-10} in the transmission channel from the serialiser board to the deserialiser board. This is highly sufficient to ensure error free digital transmission due to the error correcting capabilities of the FEC. The improvements by the FEC correspond to a coding gain of 6 dB, which in the experiments should be compared with the difference in the transmission properties from 10.0 Gbit/s to 10.7 Gbit/s.

As the 10G Home box resides at the end-user, special safety considerations must be introduced. Thus, laser equipment is located in so-called unrestricted area, where it is possible for an uneducated user to get access to the fibres and connectors. Therefore, the transmission system has to comply with the safety regulations for operation of laser products as defined in the

European Standard [85]. To avoid costly power-down equipment it is desired to keep the safety level within hazard level 1, which specifies that the total power for the 15xx nm and the 13xx region should be below 10 mW and 8.85 mW, respectively. Taking the lowest value, this equals a total optical power of 9.5 dBm. Hence, an output level for each laser below 4.7 dBm is sufficient to comply with the safety hazard level 1, why neither labelling nor power down equipment is required.

The power consumption from several racks of laser equipment at the central will be a point to consider, as the cooling of the central is costly and power consuming. Hence, the power usage should potentially be reduced as much as possible. This reduction of the power usage is as well an indication that uncooled optical devices are desired as forced cooling is power consuming and expensive.

5.3.1.2. Integration requirements

As shown in Fig. 5.8, one objective is to integrate more or less all the electronics into one CMOS chip and the optical lasers, receivers and multiplexers into one InP chip. The integration of the optical component raises some problems. Firstly, all the components should preferably be fabricated of the same material, which might constrain the choice and flexibility of components. Secondly, the main problem for integration today is the low yield when producing optical components. Typically the yield of a single laser is 30%, i.e., 70% of the devices fail due to general failure or because they do not satisfy the specified requirements regarding, e.g., wavelength stability, output power level etc.

The approach to improve yield is to relax the requirements to the components. This could be accomplished by allowing larger parametric variations, e.g., wavelength stability toward variations in temperature. In addition, keeping the components simple usually improves the yield. This is the driving force for choosing directly modulated lasers.

From a system perspective, the use of cooling of the individual lasers is undesired due to power consumption. From an integration perspective the desire to avoid Peltier elements for active cooling is backed, as Peltier elements cannot easily be integrated on the same substrate as the lasers. Hence, active cooling should be avoided if at all possible.

5.3.2. Transmission system design

The guidelines presented in the previous section are used in the following to propose a recommended solution. However, due to component unavailability it was not possible to implement the recommended solution, why an

alternative solution is suggested and implemented, which still validates the concept of the recommended solution.

5.3.2.1. Recommended solution

The most obvious choice is to use one wavelength for each channel spaced according to the wavelength spacing specified by CWDM [78]. Using distinct wavelengths for the different channels, i.e., four wavelengths in total, make reflections in connectors, splices etc. insignificant as they are filtered in the CWDM devices. In addition, the insertion losses by the passive components for a complete wavelength separated solution are acceptable. The improved insertion loss of the four-channel solution can be used to lower the optical output power with 2-4 dB, which reduces the required power to the laser, while relaxing requirements to the laser drivers.

In Fig. 5.9 the recommended architecture with four wavelengths is sketched including the laser sources, the multiplexer, transmission fibre, the demultiplexer and receivers.

According to [78] the spacing between the wavelengths should be 20 nm with wavelength sets as the following examples:

In the 13xx nm region: 1270, 1290, 1310 and 1330

In the 15xx nm region: 1530, 1550, 1570 and 1590

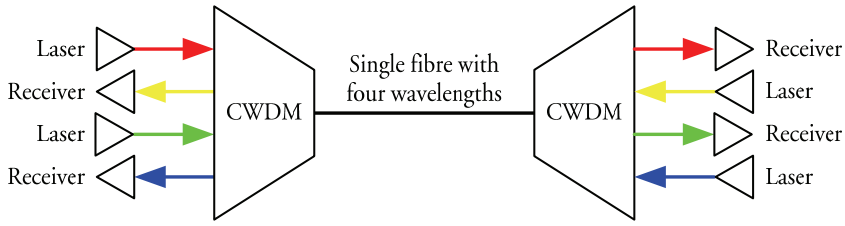


Fig. 5.9: The recommended architecture. Four wavelengths are utilised and CWDM devices are included to multiplex and demultiplex the channels

Only lasers from one wavelength region should be integrated on a single substrate as this improves the flexibility with respect to optimising the material for the wavelength. However, lasers from different regions can be used in the end terminal at the central office and at the home user.

5.3.2.2. Implemented solution

As indicated above, it was not possible to obtain lasers with direct modulation for four different wavelengths. Actually, in the 13xx nm wavelength re-

gion only lasers operating at 1310 nm were obtained, and in the 15xx nm region, only lasers operating at 1550 nm were available.

In order to allow four logically separate 10 Gbit/s channels with only two wavelengths, it is necessary to separate the forward and backward travelling signal as shown in Fig. 5.10.

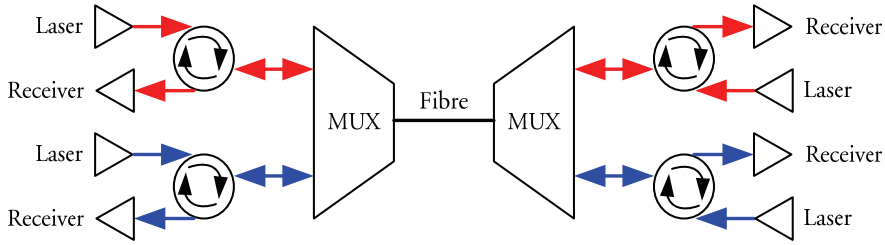


Fig. 5.10: Implemented solution based on two wavelengths and circulators to separate forward and backward travelling signal.

Here, the red laser transmits light to the circulator, which passes the signal to the waveband multiplexer. At the receiver the circulator ensures that the incoming signal is transmitted to the receiver. The main issue in this setup is to avoid any reflection from the near-end laser to disturb the signal at the receiver.

For full specification of the 1310 nm and the 1550 nm device the reader should consult the datasheets in [86] and [76].

5.3.2.3. Power budget

The power budget for the two solutions determines the required launch power. In Table 5.3, the loss parameters for all required components for both the recommended and the implemented solution is provided, where it is noted that conservative values are used. In addition, the transmission losses for standard single mode fibre are indicated.

A number of parameters are similar in the two approaches; however, one should keep in mind that the fibre loss for 1310 nm is higher than for the 1550 nm channel.

Table 5.3: Loss parameters for the passive optical components including fibre

Component	Loss pr. unit	Comments
CWDM	1 dB	Typical test sheet: 0.57 – 0.63 dB
Circulator	1 dB	Typical test sheet: 0.53 dB
WDM band splitter	1 dB	Typical test sheet: 0.3 – 0.8 dB
Splice	0.1 dB	Normally estimated to 0.02 – 0.07 dB
Fibre loss @ 1310 nm	0.4 dB/km	0.4 pr. km according to ITU-T G.652 [77].
Fibre loss @ 1550 nm	0.25 dB/km	0.35 dB/km according to ITU-T G.652[77]. 0.2 dB/km is, however, de-facto standard
Clean SC/PC connector	0.5 dB	0.2 – 0.4 d B normal

With these figures the channel loss is calculated for the recommended and implemented solution, and it is assumed that a loss corresponding to six connectors are present "somewhere" in the transmission path. In the recommended solution the lasers are connected directly to the multiplexing device through a splice, however, in the implemented architecture it was chosen to use connectors to increase flexibility. Hence, in the recommended solution there are one splice and one connector in each end totalling two splices and eight connectors. The fibre loss is based on transmission in the 13xx nm wavelength region

$$\begin{aligned} C_{L,recommended} &= 0.4dB/km \cdot 10km + 2 \cdot 1dB \\ &= 6.0dB \end{aligned} \tag{5.5}$$

The channel loss for the implemented structure is calculated for the 13xx nm channel with the highest loss:

$$\begin{aligned} C_{L,implemented} &= 0.4dB/km \cdot 10km + 4 \cdot 1dB \\ &= 8.0dB \end{aligned} \tag{5.6}$$

Note that the loss is reduced with 1.5 dB in case 1550 nm lasers are used.

The real sensitivity of the receiver is not only dependent on the photo diode sensitivity as the succeeding amplifiers should also be able to amplify the signal. With the photo-diode and TIA combination used in the implemented solution, a real sensitivity of -15.8 dB was obtained. Including 5 dB loss for connectors and splices this provides acceptable launch powers of -4.8 dB and -2.8 dB for the recommended and implemented solutions, respectively.

5.3.3. System integration

The circulator solution was implemented in a 1U rack mount as shown in Fig. 5.11. The 1550 nm directly modulated laser (red circle) was equipped with temperature control and bias circuitry, while the remaining components were delivered on EV-boards with such functions included.

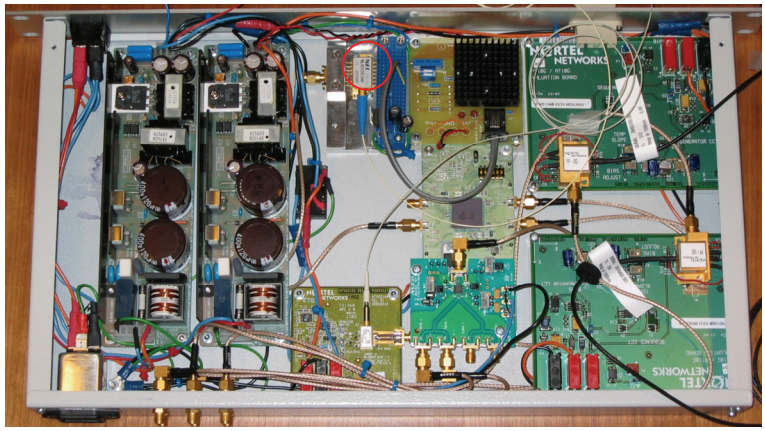


Fig. 5.11: Picture of the optical front-end in a 1U rack box. The passive optics is not shown in the picture.

In addition to the optoelectronic components, power supplies for a number of voltage levels are included in the rack.

5.4. Performance evaluation

The implemented solution for a 2x bidirectional 10 Gbit/s point to point FTTH system comprises directly modulated lasers of 1310 nm and 1550 nm. Although this laser configuration is not the recommended, the following section addresses the performance of these lasers and their integration with the FTTH digital electronics.

The main objective of the verification is to evaluate and indicate whether the chosen laser types are candidates for simple lasers in future high speed FTTH networks given the constraints described in section 5.3.1.

First the different verification scenarios are described in section 5.4.1. Then in sections 5.4.2 and 5.4.3, the two lasers in the project are evaluated separately.

5.4.1. Verification scenarios

Focusing on the lasers in the implemented FTTH solution, this section outlines the different verification scenarios for identifying whether the lasers are suitable for the job. First the transmission capabilities of the lasers are evaluated through measurement of BER in addition to inspection of eye-diagrams. Secondly, the integration with the digital part of the FTTH electronics provided by an external partner is evaluated and finally, the sensitivity of the system towards variations in the temperature and bias current is investigated. This indicates whether the CWDM technology and un-cooled devices for 10 Gbit/s operations can be combined.

5.4.1.1. Transmission scenario

The system specifications require a transmission range of at least 10 km. In the following the ability to reach this and further distances are evaluated.

A perfect input signal to the box was ensured from a laboratory pattern generator operating at 10.7 Gbit/s. Then, the signal was amplified and transmitted through different fibre spans, received and sent to an oscilloscope and a BER tester.

Two main limitations apply for transmission distance for the 1550 nm and the 1310 nm system, power and dispersion.

The fibre loss for the 1550 nm system is inherently the lower of the two, but the dispersion will apply for longer distances. The study evaluates the fibre loss and the dispersion limit.

On the other hand, the 1310 nm laser is operating in the zero-dispersion area of the fibre, why the dispersion is negligible, but the power loss is higher. Here, the study focuses on the power loss.

The length of the available laboratory fibre spans are 1 km, 2 km, 3km and 4 km. To avoid impacts from bad connectors or any other misbehaviour from the spans it was chosen never to replace fibre when increasing the fibre length. Lengths of 3 km, 6 km and 10 km were used in the measurements and they were thus connected as following:

- Fibre length of 3 km: 1 km + 2 km
- Fibre length of 6 km: 1 km + 2 km + 3 km
- Fibre length of 10 km: 1 km + 2 km + 3 km + 4 km

The only exception from this is during the measurements of the dispersion in section 5.4.3.1, where a single fibre span of 25 km is used.

5.4.1.2. Integration scenario

In order to verify the system interface to the FTTH electronics, one of the optical front-ends systems was installed to establish optical communication between two sets of FTTH terminals as shown in Fig. 5.12. The 1550 nm and the 1310 nm channel was tested separately with fibre lengths of 0, 3, 6 and 10 km

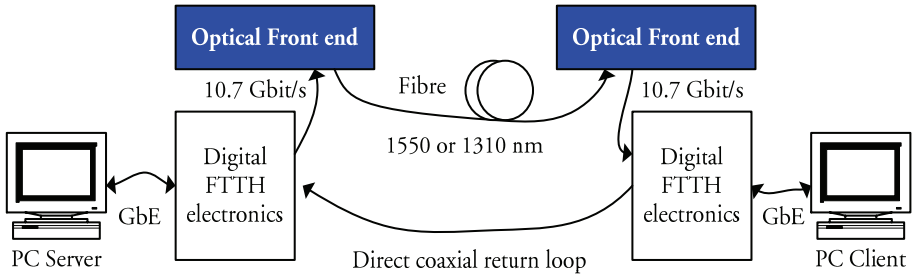


Fig. 5.12: Integration test with digital FTTH backends and implemented optical front ends

The integration of the optical front ends and lasers with the remaining FTTH electronics is verified from a PC-to-PC perspective. Every channel is addressed independently, why the return path uses a direct coaxial cable.

A PC operating as server is equipped with Gigabit Ethernet interface attached to the FTTH electronics box. The differential output of the digital box is then attached directly to the modulator driver which drives both the 1310 nm [86] and the 1550 nm laser. The optical output is launched through the fibre loop and into the receiver. The electrical output of the receiver is sent to the digital electronics and the client. The return path is formed simply by a coaxial cable.

As no separate clock signal is available, it is not possible to obtain any eye-diagrams or BER measurements. Instead a simple procedure was established:

- The client is programmed to transmit ping requests
- Bias current of lasers is slowly decreased until ICMP transfer breaks
- A multimedia application is active without buffer to see effect of packet loss close to the limit.

Even though this is far from a perfect scientific approach it is considered sufficient to give an indication of the operation performance.

5.4.1.3. Wavelength stability

In the recommended solution, four wavelengths in the same wavelength region are suggested. This sets requirements to the wavelength stability, as the variations should fit within the standardised pass band of the CWDM channels.

Furthermore, wavelength stability is usually ensured by controlling the temperature. However, due to power and integration concerns, it is desired to avoid the temperature control, why it should be verified that the removal of the control will not alter the wavelength more than allowed by the spectral width of the pass band. In the wavelength stability tests the lasers are evaluated with alterations of both bias current and temperature.

5.4.2. Directly Modulated 1310 nm laser

5.4.2.1. Transmission capabilities of 1310 nm laser

The eye-diagrams after the limiting amplifier at the receiver were recorded with the different fibre spans as shown in Fig. 5.13.

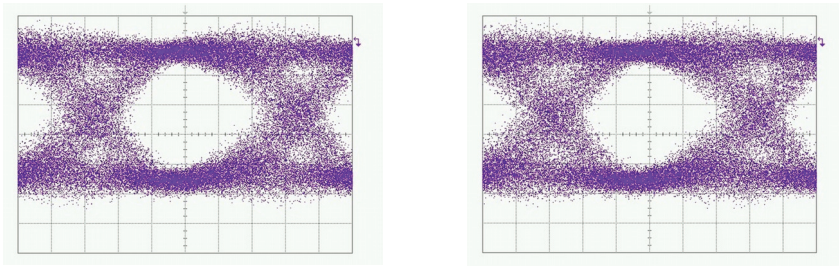


Fig. 5.13: Eye-diagrams of the received electrical signal after the limiting amplifier for back to back and 10 km fibre span. The time scaling is 15 ps/div and vertically it is 100 mV/div. An extinction ratio of 9-10 dB is observed. The main noise contribution is thermal noise due to now input current to the scope.

The quality of the eye-diagrams is quite similar and well-defined eye-openings are observed. It is noted that the limiting amplifier performs 2R regeneration of the signal, which influences the quality of the eye-diagrams. The eye-diagrams for fibre spans of 3 km and 6 km showed similar shape and quality.

The performance of the 1310 nm laser and the associated channel was further evaluated by BER measurements for the fibre spans of 3, 6 and 10 km. The results are shown in Fig. 5.14.

As expected the effect of dispersion is completely absent and there is no evidence of any distortion of the signal with the fibre length. It is thus indicated that the maximum fibre length is dependent only of the launched power as long as higher order non-linear effects are absent. These could otherwise arise with very high launch powers as self-phase modulation and four-wave mixing.

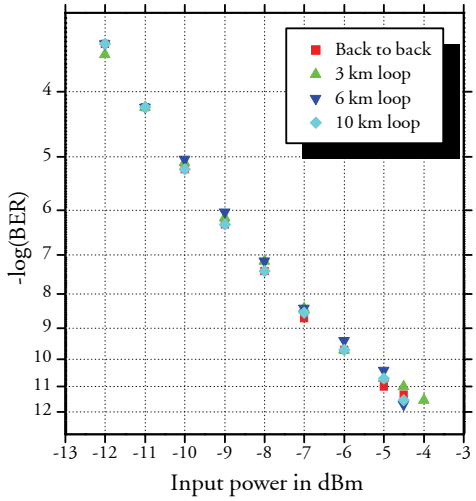


Fig. 5.14: BER measurements for the 1310 nm laser obtained after the fibre. No power-penalty due to signal distortion observed. It is noted that no optical preamplification is present.

5.4.2.2. Integration with FTTH electronics

The 1310 nm laser channel was integrated with the digital FTTH electronics delivered by the external partner. For fibre lengths of 0, 3, 6 and 10 km the output power level and the power level at the receiver was monitored to indicate the power level of connection failure.

Two main limitations are observed from the table. For short fibre spans the limiting factor is the launch power, and for longer fibre spans and higher attenuation the received power is limiting.

Table 5.4: Power levels when connection between boxes fails for the 1310 channel.

Fibre length	Launch power	Power at receiver
0 km	-4,3 dBm	-4,3 dBm
3 km	-3,4 dBm	-5,7 dBm
6 km	-2,7 dBm	-7,4 dBm
10 km	-2,7 dBm	-9,6 dBm
10 km + 5 dB attenuation	-2,5 dBm	-14,3 dBm
10 km + 10 dB attenuation	n/a	n/a

The first limitation is caused by the operating conditions of the laser, which provides an inferior signal quality if the bias current is too low, i.e., the working point goes below the linear slope of the P/I characteristic of the laser.

The latter limitation for longer fibre spans are caused by the sensitivity of the receiver and as explained in section 5.3.2.3, a better sensitivity than -15 dBm is hardly expected. Whenever the system looses the link, it is usually required to increase the power of 1-2 dB to re-establish the link.

5.4.2.3. Wavelength stability

The 1310 nm laser is not actively temperature regulated, why the observed effects may arise from both the current induced and the heat induced wavelength variations. The optical spectra for the 1310 nm laser with output power of -3 dBm and +3 dBm is shown in Fig. 5.15.

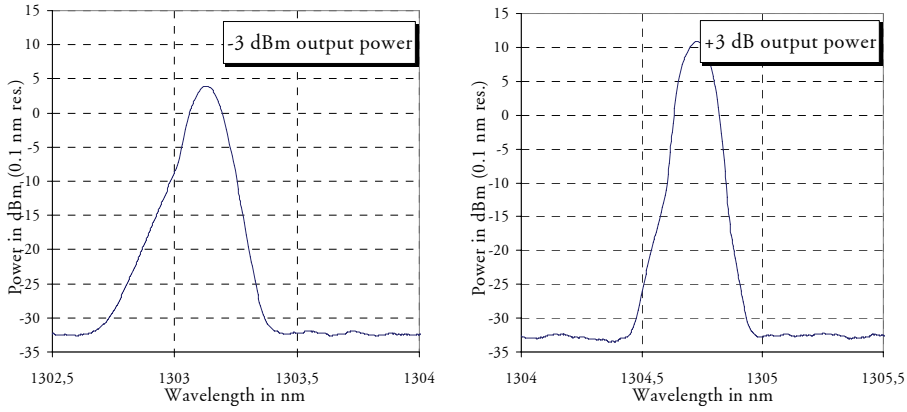


Fig. 5.15: Optical spectrum of 1310 nm laser. A shift from 1303.1 to 1304.2 nm is observed

A wavelength shift of 1.1 nm is observed and the spectrum at low bias current is significantly broader than for higher bias current. Again, this is probably caused by too high modulation power in connection with too low bias current. It is also noted that the centre wavelength of the device is quite far from the specified 1310 nm.

5.4.3. Directly Modulated 1550 nm laser

The 1550 nm DM laser was provided by NEL/NTT without temperature regulation or bias circuitry. This was added afterwards as seen in Fig. 5.11 with the red circle. The availability in 2002 of directly modulated lasers in the 1550 nm region was quite limited, as they are not suitable for long-haul communication due to the broad spectrum that disperses in the 15xx nm region.

5.4.3.1. Transmission capabilities of 1550 nm laser

As for the 1310 nm laser, the performance of the 1550 nm laser was accessed through eye-diagrams and BER measurements. First, as sketched in Fig. 5.16, the eye-diagrams for back-to-back and 10 km fibre span is observed after the electrical amplification and 2R regeneration.

The eye-opening in the two eye-diagrams are quite comparable, although the diagram for back-to-back has an eye-opening that is slightly broader. The eye-diagrams for fibre lengths of 3 and 6 km showed similar shape.

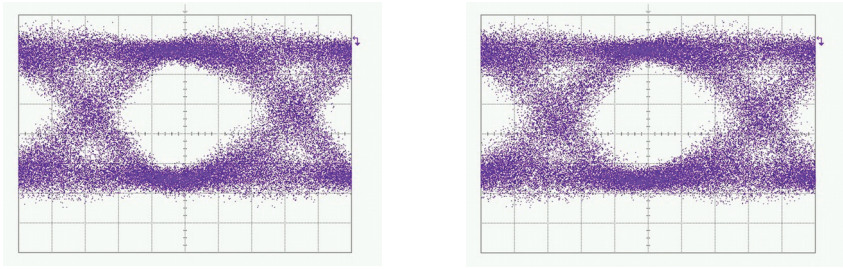


Fig. 5.16: Eye-diagrams of the received electrical signal from 1550 nm laser after the limiting amplifier. The signal corresponding to 10 km transmission is slightly noisier than for the back to back signal. The horizontal scale is 15 ps/div and vertically it is 100 mV/div.

The performance of the 1550 nm laser operation was further quantified by BER measurements as shown in Fig. 5.17.

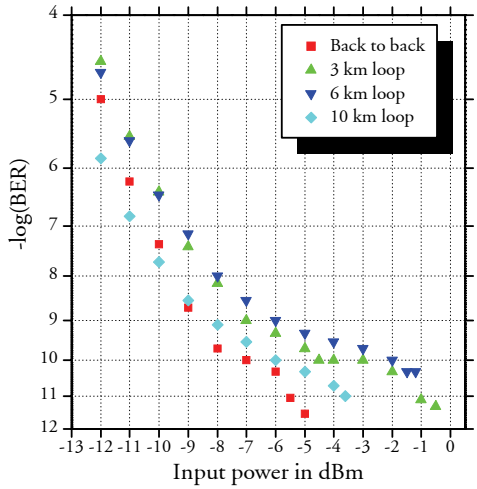


Fig. 5.17: BER measurement of 1550 nm laser with different lengths of fibre spans.

As no clock recovery was present the slowly varying delay in the fibre spans were compensated by manually adjusting the delay of the BER tester while observing the delay on the scope. This method was especially used for measurements of BER below 10^{-11}. The measurements were obtained by attenuating the optical power before the receiver with an attenuator, i.e., the bias level and the output power of the lasers were unaltered. The manual clock phase adjustment is believed to cause part of the inferior measurement.

The measured receiver sensitivity is close to -9 dBm (BER @ 10^{-9}), and a power penalty of approximately 3 dB is observed for the 6 km loop. The back-to-back BER curve shows a deviation from a straight line and the measurement with the fibre span of 10 km shows significantly better results than that of 3 and 6 km. Whether this might be caused be "optimal" deformation of the signal is unknown. The floor tendencies can be explained both by dispersion limitations and by missing clock recovery, which for small error rates makes the measurement very sensitive to even small temperature variations during the measurements. To further evaluate the dispersion effects for the directly modulated 1550 nm laser a special setup was established including optical amplification before the receiver. This enables investigation through far longer fibre spans; however noise from the amplifier is added. During these measurements there was no access to equipment to set up a full BER measurement, why only eye-diagrams are used in the evaluation.

The optical eye-diagrams after fibre spans of 0 km, 10 km, 25 km and 35 km are shown in Fig. 5.18.

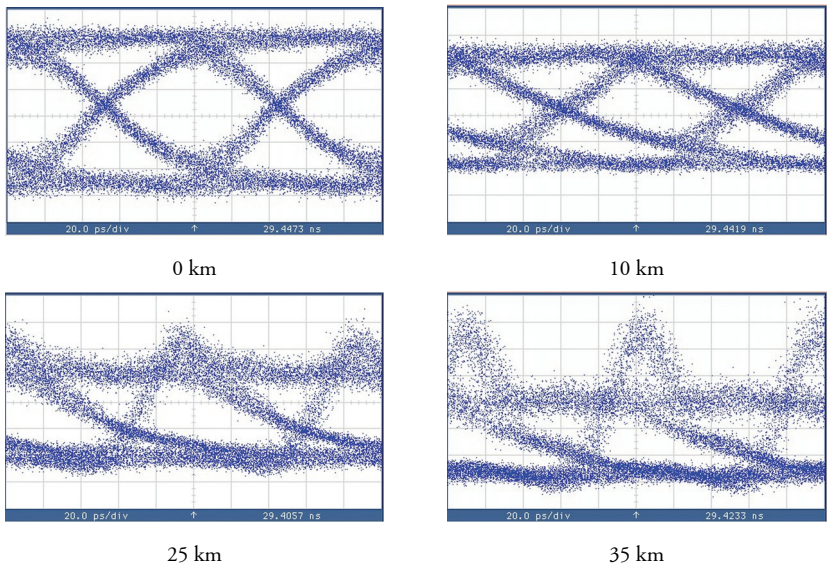


Fig. 5.18: Eye-diagrams of signal from 1550 nm laser recorded in the optical domain. Optical amplification is used to isolate the effect of dispersion.

The fibre span of 10 km is formed as described in the beginning of this section, the 25 km is a single fibre span and the 35 km is the sum of all the fibre spans. The effect of the dispersion is that some of the spectral components in the modulated signal have a different group velocity than others. The directly modulated signal is spectrally broad, why different frequencies in the signal travels with different speed causing pulse dispersion.

With increasing length it is observed that the rise and fall times differ, why the eye-opening becomes rectangular, which is clearly seen for the eye-diagram at 10 km. For longer spans the rise time is reduced, which result in overshoot of some of the pulses when going from space to mark.

In summary, the eye-diagrams are well defined for fibre spans up to 10 km and even for 25 km a clear eye-opening is observed. For 35 km the performance is questionable. It is, however, obvious that splices should be preferred compared to connectors whenever possible due to the significantly lower loss value.

5.4.3.2. Integration with electronics

The 1550 nm laser was integrated with the FTTH digital electronics to verify the performance. As for the 1310 nm laser, fibre spans of of 0, 3, 6, and 10 km are used, and for each fibre length the output and the received power

was monitored when the communication failed. The results are given in Table 5.5.

The results are quite similar to those obtained for the 1310 nm channel. Two limitations apply; the launch power for short fibre spans and the receiver sensitivity for long spans. It is thus demonstrated that the system operates with the 1550 nm laser through 10 km fibre with 10 dB margins. When the system looses the link it is usually required to add 1-2 dB optical power to re-establish the link.

Table 5.5: Power levels when connection fails for the 1550 nm channel

Fibre length	Launch power	Power at receiver
0 km	-2.9 dBm	-2.9 dBm
3 km	-1.5 dBm	-4.7 dBm
6 km	-1.2 dBm	-5.6 dBm
10 km	-0.8 dBm	-7.2 dBm
10 km + 5 dB attenuation	-0.7 dBm	-12.0 dBm
10 km + 10 dB attenuation	2.1 dBm	-14.6 dBm

5.4.3.3. Wavelength stability

The wavelength dependency of the 1550 nm NTT laser was evaluated by observing the spectrum with different bias levels. The result, when using -3 dBm and +3 dBm output powers are sketched in Fig. 5.19.

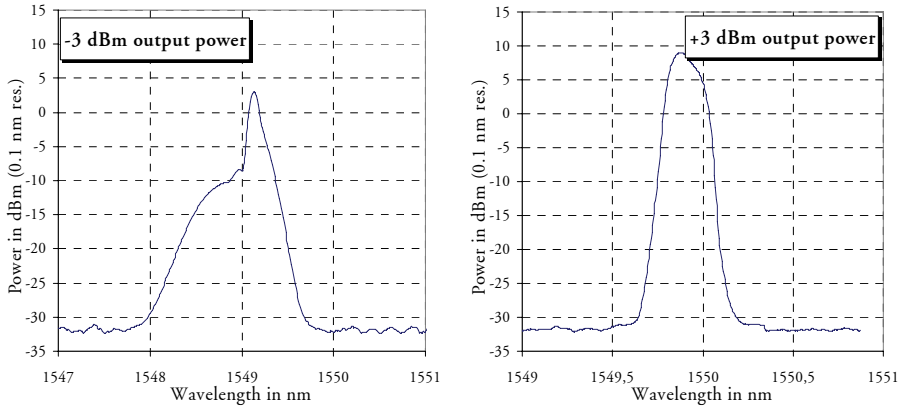


Fig. 5.19: Optical spectrum of 1550 nm laser with an output power of -3 and +3 dBm. A shift from 1549.2 nm to 1549.7 nm is observed with an output power of -3 and +3 dBm. A shift from 1549.2 nm to 1549.7 nm is observed.

It is seen that the wavelength is only slightly dependent on the optical output power, as the wavelength shift is only close to 0.5 nm. The shape of the spectrum with small current is extremely broad and not symmetric, which indicates non-optimal operation at this bias current with the applied modulation current. Usually an increase in the bias current heats up the laser cavity and the wavelength is altered; for the 1550 nm laser this effect is compensated by the temperature control.

It was possible to verify the wavelength stability of the 1550 nm laser towards the temperature as this comprise a Peltier element for cooling or heating. The implemented temperature control is only designed to cool the device, as the opposite is rarely required. The bias current was set for maximum output power, and the temperature was thus regulated from 0° C to 45° C and the optical spectrum was monitored as reported in Fig. 5.20.

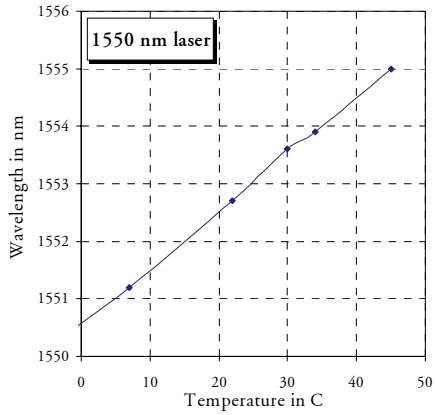


Fig. 5.20: Dependency of 1550 nm laser wavelength towards temperature.

It is observed that the wavelength over temperature ratio forms a linear function with a span of 4.5 nm within the observed temperature interval. Extrapolating the range to 0° C to 70° C gives a wavelength span up to 7 nm. It is noted that for temperatures above 40° C the effect of the temperature controller is totally absent, as the heat sink and the ambient temperature was sufficient to keep the temperature down. Hence, if a span of 7 nm is allowed, the 1550 nm laser is able to operate without any active temperature regulation. This underlines and supports the use of CWDM devices with passbands of 13 nm in the recommended solution.

5.5. Discussion

The performed case study is used to indicate answers to some questions that were raised prior to the project. This section describes these answers and suggestions through a discussion of the findings of the theoretical and practical evaluation.

5.5.1. Lasers

One main issue is to determine whether direct modulation is suitable for 10.7 Gbit/s transmissions in the distances to be considered within FTTH, i.e., distances around 10 km. Initially, a system design with lasers in the 13xx nm wavelength region is preferred as they seemed more available. However, as this was only the case for 1310 nm, it was necessary to include a 1550 nm directly modulated laser.

Directly comparison of the 1310 and the 1550 nm laser will not necessarily provide any fruitful result as it is hard to separate inherent properties for the two wavelengths and the differences in the laser design for the actual components. With this in mind, it is without doubt that the 1550 nm laser did perform best of the two with highly sufficient launch power.

The main problem with the 15xx nm region is the dispersion, which cancels standard transmission with directly modulated lasers for distances longer than 15-20 km depending on the fibre and the spectral broadness. If such distances are accepted, the 15xx region has the advantage of lower fibre loss, which directly translates to potentially lower output power.

For both lasers the temperature variations were no problem, as the related wavelength variation in all cases were within the pass band of the CWDM channels.

Other studies [79] has shown acceptable performance of 1310 nm lasers with even very high temperatures, which indicates a slight preference for lasers in the 13xx nm region.

The costs of the components today also favours the 13xx wavelength region, however, it is believed that price reflects the produced quantity rather than the actual cost if real mass-production is initiated. E.g., due to the dispersion, there has been no market for long-haul DM 15xx lasers, why the cost today is very high.

Hence, the properties within the wavelength region towards monolithic integration should be the main decision point, whether lasers in the 13xx nm, 15xx nm or both wavelength regions should be used for FTTH applications.

5.5.2. Receivers

In the experimental verification performed in this study, only PIN photodiodes integrated with transimpedance amplifiers were assessed. The limiting factor for the receiver sensitivity in the implemented solution was the amplification in the TIA rather than the optical sensitivity of the photo diode.

Although the sensitivity with APD receivers is improved it is questionable whether the improved performance outweighs the more complicated design.

5.5.3. System architecture

It is recommended to implement the four channel system on four separate wavelengths in either the 15xx nm or the 13xx nm region. The wavelengths are multiplexed by CWDM wavelength multiplexers with a channel spacing of 20 nm. This allows for operation of uncooled lasers, where wavelength fluctuations may apply. Furthermore, the passive optical equipment only imposes a few dB insertion losses in the optical channel.

It is not recommended to use the two wavelengths solution, which is implemented as part of this study. While it did operate as expected, care should be taken when connecting fibres to avoid any reflections, which easily may be disastrous. Furthermore, the inclusion of circulators increases the insertion loss and the use of lasers in two different wavelength regions complicates the monolithic integration of the lasers and the receivers.

5.6. Summary

Today, the last mile between the service provider point of presence and the end user is dominated by copper technologies as ADSL and cable modems. The reach of the DSL technologies is very dependent on the length of the access loop, and in the best cases bit rates up to 24 Mbit/s for ADSL and 200 Mbit/s for VDSL connections can be expected; the latter only for short distances up to a couple of hundred meters.

While these bit rates are sufficient for web browsing, email etc., they are insufficient for future multimedia services like, e.g., HDTV, VoD and high quality video conferencing. Fibre access is necessary. The main challenge for deploying fibre access for the last mile is the user's willingness to pay a lot more for the significantly increased capacity. Compared to a telephone conversation, the user is probably only be willing to pay 1:10000 pr. bit for new

multimedia services, hence the paradigm of paying for bits is exchanged with payment pr. service.

In this chapter, the selection of optical and electronic components for future high speed FTTH solutions have been evaluated based on a case study. The case study comprised two bidirectional point to point channels of each 10 Gbit/s over a single fibre infrastructure. A distance of 10 km was considered and the system should interface to 10.7 Gbit/s digital CMOS electronics. The objective of the case study is to evaluate the performance of simple components like directly modulated lasers in future fibre access networks.

The laser evaluation included a 1310 and a 1550 nm laser, which were both directly modulated. This was chosen, as directly modulated lasers with their simple design are more obvious candidates for integration and cost-reduction compared to externally modulated lasers. On the contrary the directly modulated lasers provide a broader spectrum of the emitted light, which in the 15xx nm region translates to dispersion.

It is concluded that both lasers operated with the digital FTTH electronics over the considered distances. The dispersion for the 1550 nm laser was further studied, and it is indicated that for the specific laser the dispersion effect was only significant for distances above 25 km. Both lasers were accessed to verify their performance in non-temperature regulated environments. A wavelength drift of up to 7 nm for a temperature variation of 70°C was extrapolated from the measurements. This is within the passband of CWDM devices and hence not a problem for use in FTTH systems.

Based on the evaluations there are no clear indications whether 13xx or 15xx lasers are most suited for FTTH applications. However, other evaluations in the literature indicate that 13xx lasers are less sensitive to very high temperatures.

The main result of the study indicates that directly modulated lasers at 10.7 Gbit/s are viable candidates for future FTTH access networks. Both the trialled laser at 1550 and 1310 operated satisfactorily at distances up to 10 km, and it was indicated that the dispersion limitations for the 1550 nm was above 25 km. The choice between the two wavebands depends on the system design and whether amplification should be required, which is most mature in the 15xx nm waveband due to EDFAs.

The viability of the simpler directly modulated lasers allows more cost-efficient integration with higher yield, which will make the FTTH equipment and deployment more attractive for the very cost-focused customer.

6. Dynamic Bandwidth Allocation

The user perceived quality of a network is dependent on its capability of providing sufficient services to "its" applications. While this is evident, it is often in direct conflict with the demand to use the network capacity as efficiently as possible at all times. Attempts to reduce the delay by guaranteeing bandwidth also reduces the utilisation of the network as the reservation of capacity inherently results in an under-utilised network.

The utilisation in completely packet switched networks is very high and often the capability to transport delay sensitive applications has been introduced by implementing a huge on-average over-provisioning. However, the burst oriented traffic characteristics of many TCP/IP based applications result in a self similar traffic pattern even at high bandwidth scale [87]. Thus, for applications with high demands on bandwidth *and* delay, such over provisioning does not provide a solution to the problem.

In the other end of the scale, the fully circuit switched nature of the telecommunication networks provides very well defined channel characteristics on the expense of a lower utilisation. This is, apart from expensive, not useful for many data applications.

While the ASON framework, as described in section 2.3 provides an interoperable control plane to horizontally integrate heterogeneous networks, no standardised approaches exist for the vertical integration of the applications and the network control plane. The work described in this chapter suggests an approach for communicating application resource requests to an interoperable network control plane. The work is closely related to the ongoing European IST project MUPBED, why parts of the implementations are "in progress". The general objectives and results of the MUPBED project are described in section 6.1 followed by more general considerations on the vertical integration of a dynamic transport network in section 6.2. In section 6.3, algorithms for implementing application driven resource triggering are provided including identification of parameters to control the level of dynamics in the network. The discussion on the level of dynamics is further elaborated in section 6.4, which describes modelling activities that are initiated to investigate lower limits of the dynamicity. Finally, in section 6.5, a

summary of the work is given including an outlook of the future perspectives within vertical service integration in heterogeneous networks.

6.1. European IST project MUPBED

Today, the transport networks are based on a fixed set of links using, e.g., SDH/SONET transport technologies on top of optical wavelengths. The management of these optical wavelengths is far from dynamic and connection placements are done manually.

The European IST project MUPBED[1] addresses this by demonstrating scalable, flexible and dynamic transport network infrastructures for research applications. A detailed motivation for the project is provided in this section highlighting the importance of switched connections and integration of applications. Final results of the project are not included, as the MUPBED project is ongoing until summer 2007; however, a few expected results are provided.

6.1.1. Motivation and main objectives

In today's networks, there is a complete decoupling between the networks and the applications utilising the network, i.e., the transport network is statically configured and does only consider the transportation of bits regardless of which application they represent. On the other hand, the applications only consider the IP layer, thus they do not care about the underlying transport network. This decoupling is convenient as the network operator only needs to consider the network and the application user only needs knowledge of the IP layer. The drawbacks, however, are networks not really adapting to the high demands of future applications, and networks with inferior utilisation, because considerable over-provisioning is introduced to tackle the bursty nature of the data-traffic. To improve this, a more holistic network view is required and it is necessary to consider, how high demanding applications can communicate their requirements to the network in order to get the requested service level. Furthermore, the utilisation of the network resources should be improved by introducing switched connections through an interoperable control plane. These two integration issues are illustrated in Fig. 6.1.

[1] MUPBED: Multi-Partner European Test Beds for Research Networking. The project started July 2004 and is expected to end by June 2007

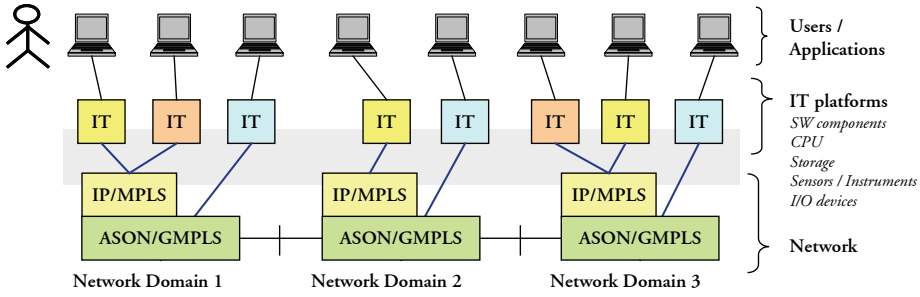


Fig. 6.1: Vertical and horizontal integration in MUPBED [88].

Vertically, the users or applications communicate through specific functions with the network control plane in order to request sufficient resources for specific connections. Horizontally, the different heterogeneous network domains are integrated by an interoperable control plane. Hence, several objectives motivate the MUPBED project, although they can be boiled down to one single word, "integration". In addition to the vertical and horizontal integration described above, it is within the scope of the project to increase a common understanding between network operators and application users of which functions belong to which part in the network.

It is clearly an objective to use standardised interfaces to interconnect networks, why collaboration with standardisation bodies like OIF, ITU-T and IETF is part of the project. Hence, it is expected that the outcomes of the project will influence the standardisation.

The MUPBED consortium comprises partners from the telecom operators, national research and education networks (NREN), academia, equipment providers and application users. In total 16 partners from 8 different European countries.

6.1.2. MUPBED test bed

The heart of the MUPBED project is the large scale European experimental test bed as shown in Fig. 6.2. This is used mainly for the horizontal integration and verification. The MUPBED network comprises five test beds in which all the switching functionalities reside. These test beds are interconnected by layer 2 connections through the NRENs and the European GÉANT network [89].

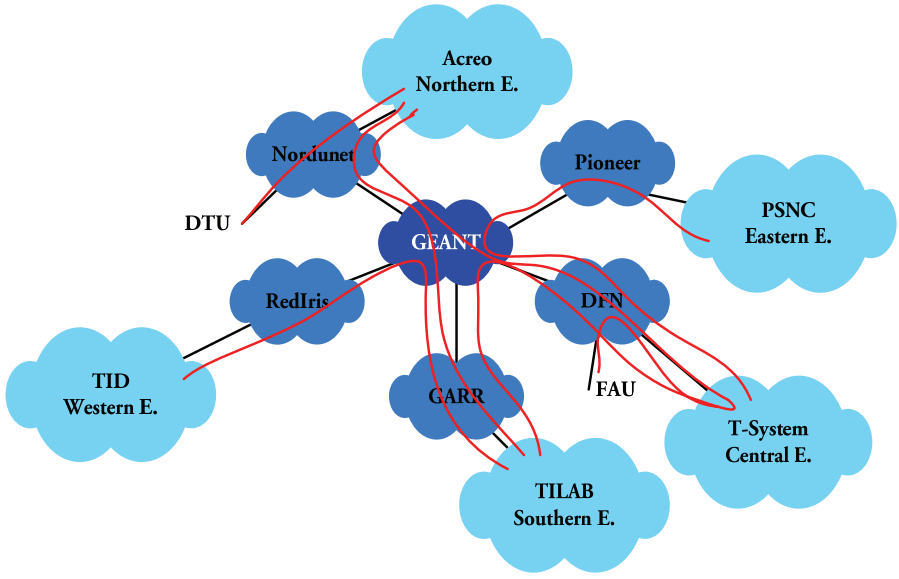


Fig. 6.2: MUPBED test bed and layer 2 connections. Test beds in light blue, NRENs in blue and GÉANT in dark blue.

The research institutions, e.g., DTU and FAU[1] are connected to the MUPBED network through one of the test beds. If DTU requests a connection to FAU, this will pass the link to the Northern European Testbed, back through NorduNet, GÉANT and DFN to the Central European Testbe. Here the connection is switched to the link passing DFN to FAU. The interconnection is primarily using gigabit Ethernet links through the NRENs and MPLS LSPs through GÉANT.

The resulting layer 3 (IP layer) data plane between the test beds is shown in Fig. 6.3, with the different technologies. Both the Central Europe Test bed and the Southern Europe Test bed located at T-Systems[2] and TILAB[3], respectively, uses ASON/GMPLS interconnected through an OIF E-NNI. The Eastern Europe Testbed located at PSNC[4] in Poland uses Ethernet technology and connects to the ASON core with an OIF UNI2.0 interface. The same interface is used to interconnect the ASON core with the Northern Europe GMPLS network located at Acreo in Sweden and the Western Europe MPLS Test bed at Telefónica in Spain.

[1] FAU: Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] T-System: Part of Deutsche Telekom

[3] TILAB: Telecom Italia research facilities

[4] PSNC: Poznan Supercomputing and Networking Centre

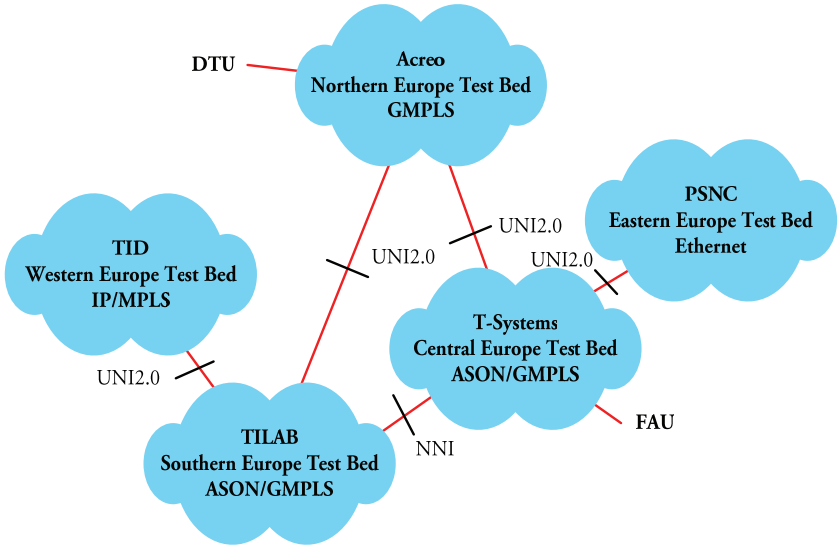


Fig. 6.3: IP layer data plane of the MUPBED test bed with reference points

It might seem strange that UNIs are implemented in the network between network domains. The reason for this is non-interoperable implementations between vendors at the E-NNI level, which is an ongoing topic within OIF and IETF.

The implementation of the test bed has already resulted in novel interoperability tests during public demonstrations in Turin and Bordeaux [Publ. 15][Publ. 12].

6.1.3. Applications to trial in the test bed

The MUPBED project focuses on the high demanding research applications existing today and in the future. While it is impossible to verify the proposed dynamic services with all possible applications, some have been selected as it is believed they represent some of the requirements to the network otherwise expected in future research applications.

The four main applications selected for implementation in the MUPBED test bed are among others:

- High quality uncompressed video production
- Content and storage
- Video conferencing

The "killer" application is obviously the high quality video production. Consider a distributed real time video production, where remote studios

require time sensitive communication with the online editing location. Compression is not possible, as this would add too much delay, why bandwidth up to 1.5 Gbit/s is required. The requirements to delay, jitter, bandwidth and packet loss is sketched in Table 6.1 for some of the pre selected MUPBED applications.

Table 6.1: Service requirements for the MUPBED applications [90]

Application	Traffic Type	Quality	Requirements		
			Bandwidth	Latency / Jitter	Packet loss
Storage Backup and Restore	Information and Data services (bi-directional)	High	100 Mbps – 2 Gbps	200 ms / -	< 0.1 %
Accelerated VoD Streaming	Real-time on demand (unidirectional)	High to Very High	50 – 100 Mbps	500 ms / 50 ms	< 1 %
Uncompressed high quality video transmission	Real-time interactive (bi-directional)	SD: High HD: Very High	300 Mbps – 1.5 Gbps	150 ms / 1 ms	< 1 %
Point to point conferencing	Real-time interactive (bi-directional)	High	1.5 – 20 Mbps	150 ms / 50 ms (skew: 80 ms)	1%
Multipoint videoconference	Real-time interactive (bi-directional)	High (MPEG2) Very High (No compression)	2 – 13 Mbps per partner	150 ms / 50 ms (skew: 80 ms)	1%

The interactive services usually require a delay less than 150 ms, which corresponds with the ITU-T recommendations [91]. It is also seen that the bandwidth requirements for the content and storage application are significant (up to 2 Gbit/s), however, the delay requirements are slightly relaxed.

The identification of the network requirements imposed by the applications is used to classify the MUPBED applications into service groups. These groups should, together with specific parameters, provide input to the network control layer how it should administrate the resources.

6.1.4. Expected results

There are more than 50 different national research and education networks around Europe. Each of these attempts to fulfil the needs of their custom-

ers, i.e., the researchers and high demanding application users connected to their network.

The main result of MUPBED is to provide ways to integrate these networks using standardised interfaces that will allow smooth integration of services passing several research networks and even commercial telecom networks.

Furthermore, it is expected that the experiments in the test bed will provide feedback that is used to push the further standardisation in IETF, ITU-T and OIF in a "MUPBED" direction. This utilises the strong engagement by MUPBED partners in these standardisation bodies.

6.2. Vertical integration of applications and networks

As discussed in the introduction of this chapter, no standardised approach is available for integrating applications with the network control plane for heterogeneous network domains. Hence, this section describes how the applications can request resources and how these requests ultimately results in modification in the circuit layer.

Roughly spoken, the dynamics of today's network is based on a client to human to management approach, where a client emails or call a network administrator who sets up the connection in the circuit switched core network manually. The authentication is based on, whether the network operator knows who he is speaking with, and the authorisation is based on a manually obtained agreement. Needless to say, such path establishment is not in the second or minute time scale; rather in weeks and months. If the connection is a multi-domain connection the client should contact each of the involved network administrators or the first administrator should contact the next in the connection sequence. This further complicates the connection setup and the time consumption is significant.

In this section a brief review of advance reservation and common vertical integration issues are provided. The adaptation function is introduced to provide the communication channel between the application and the network control.

6.2.1. The adaptation function

High demanding research applications usually do not communicate on a UNI level and the advantages by adding such functionality to the applications might be questionable. This is mainly because a separation of applica-

tions and network layer protocols, e.g. RSVP-TE, CR-LDP and interfaces, e.g. IETF UNI or OIF UNI is required by the application developers, who require a higher level of abstraction. Hence, an *adaptation function* (AF) is introduced as responsible for interfacing with the network control plane and for deciding when new network resources from the circuit layer should be established. The adaptation function receives resource requests from the applications and is responsible for triggering these resources in the network. In this way a decoupling between the applications and the actual transport technology is ensured.

The adaptation function includes functions for manipulating the circuit layer by requesting setup of new and release of existing connections. The manipulation can be driven by application requests or by long term monitoring of link utilisation and prediction of the most optimal usage parameters.

The location of the adaptation function in the network is illustrated in Fig. 6.4, where the link between the work station and the adaptation function is a direct link or through a client network. Then, the adaptation function controls the establishment of connections through the interoperable control plane of the heterogeneous transport network. The adaptation function does not consider the network topology between the UNI interfaces internally in the network as the AF is simply aware of the edge to edge connections that are associated with the UNI it controls. The AF logically controls the control plane and the data plane.

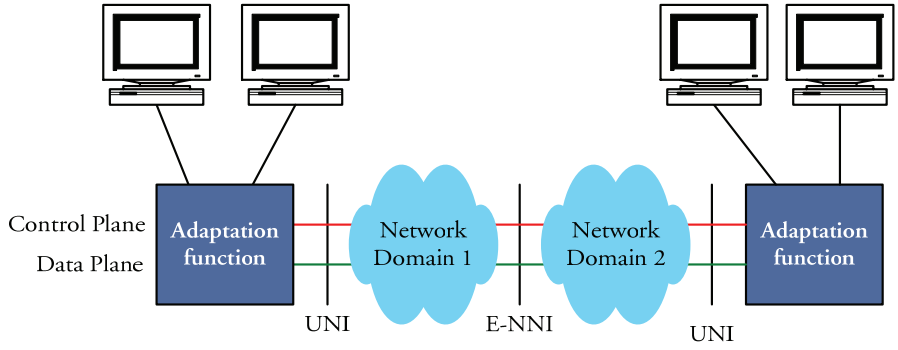


Fig. 6.4: One adaptation function serves several users for each UNI.

In addition, the figure shows how the horizontal integration of the control planes for different domains interacts with the vertical integration of the applications and the network.

A more detailed overview of the communication channel between the applications and the network is provided in Fig. 6.5, where the adaptation function comprises the functions within the dotted lines.

The applications, acting as clients, communicate with the adaptation function through an *Application Programming Interface* (API), which is implemented as a web service. The only modification at the application side is the *Network Service Requester* (NSR) component which is located in the control plane of the applications. This component communicates with the *Network Service Provider* (NSP) and the communication is based on *Simple Object Access Protocol* (SOAP), which is based on *Extensible Markup Language* (XML) technology.

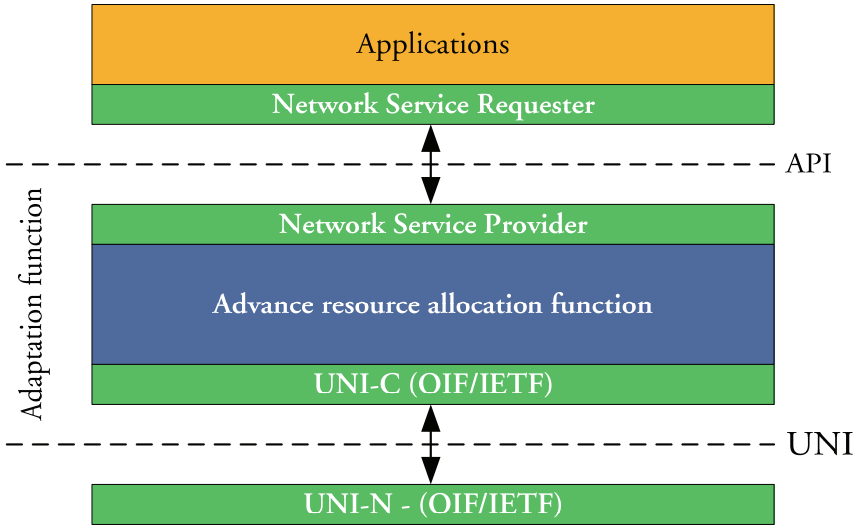


Fig. 6.5: Detailed overview of the adaptation function

The core of the adaptation function is the advance resource allocation function shown with blue in Fig. 6.5. This function is responsible for registering resource requests and allocating the resources when the connection should be established. If sufficient resources are available in the packet layers, these should be used for the allocation request. Otherwise, new resources should be allocated in the circuit layer. Providing advance reservations in the circuit layer is challenging, as this is not directly supported by the network interface. This problem might be reduced to a statistical problem by proper over-provisioning at the packet layer in combination with local control of the resource requests.

A network proxy handles the UNI to the circuit layer.

6.2.2. Advance reservation

It is desired from the application perspective to provide advance reservation, which enables the applications to reserve resources in advance. However, in order to avoid misinterpretations of "advance reservations" it is useful to define what is meant by this and related terms.

6.2.2.1. On demand reservation

An "on demand resource allocation" is defined as a reservation that is initiated in the network as soon as the request is received from the application layer. If the request is successful, the resources are allocated and the connection is only terminated if one of the following situations happens:

- Explicit release is sent from the application
- Time out due to missing refresh messages in the reservation signalling protocol (RSVP)
- Break down of the link
- Preemption initiated by the network management layer.

Thus no start and end time is provided and the connection is established and released on the fly. This is completely analogous to a plain old phone conversation, where the success of a resource allocation is a stochastic problem.

6.2.2.2. Scheduled reservation

Scheduled reservation is a reservation service, where the reservation request comes some time before the real service is required, the resource availability is verified, and the request acceptance/denial information is returned to the requestor. The actual resource reservation done at the scheduled time in the future may only fail due to network malfunction, and never due to insufficient resources.

Scheduled reservations can be made hours, days, week or month prior to the actual reservation.

6.2.2.3. Open reservation

This third definition is included as a reservation with a start time, but no end time. Hence, the resource request is stored in the network, but the reservation process is not done before the connection should be used. The termination conditions are the same as for the "on demand resource allocation". Depending on the underlying network interface, the reservation could be confirmed in advance (as in scheduled reservation) or just within a

short time period before the actual reservation time analogous to on demand resource allocation.

6.2.3. Vertical integration schemes

Although it is claimed that no standardised vertical integration has been specified, several proprietary solution for, e.g., advance reservations have been suggested. In relation to grids, it is desired to reserve slots for network resources similar to processing and storage resources. Such schemes are implemented in the GARA as part of the GLOBUS framework [92]. Other approaches for scheduling network resources in relation to grids are, among others, found in [93], [94] and [95]. More generally the Canadian NREN CANARIE has suggested the User Controlled Light Paths (UCLP) for user driven provisioning of light paths [96], and *Internet2* has suggested the Bandwidth Reservation for User Work (BRUW) [97].

Commonly for these schemes is that they are not suited for multi domain operation over OIF/IETF UNI interfaces, as the UNI do not support the exchange of routing information as is otherwise required. Hence, in the following the possibilities of providing advance reservation and vertical integration with an OIF-UNI based transport network are considered.

6.3. Resource allocation

The network type to consider in this work is based on UNI interfaces implemented with the RSVP-TE [11] or CR-LDP [10] signalling protocols. This does not allow for implementation of the suggested resource administration approaches described in section 6.2.3.

Hence, in this section it is discussed how advance reservation can anyway be emulated through the adaptation function that stores the application requests and attempts to ensure sufficient resources at the given start-time; if necessary allocations in the circuit layer are requested.

It is thus important to note that the advance reservation described in this section is basically "scheduled on demand reservations" and no guarantees for an application request is actually provided before the resources are physically allocated.

6.3.1. Resource management algorithms

The control and management of the resources in the circuit layer is controlled by three main processes, which are described in the following. The

three processes are responsible for registering resources, triggering resources and notification to the applications of the result of the resource allocation.

6.3.1.1. Request registration

The first algorithm is the resource registration function. This function is responsible for maintaining the database of application requests associated with the UNI. It is assumed that a database is available for the function. More specifically, the function should receive application requests and store these in the database. In addition, it should handle any application triggered deletion of stored application requests.

The algorithm for the function is shown in Fig. 6.6, and it is noted that the function does not actually reserve the resources. It only registers the requests in the database and returns the aggregated requests for the time period of the application request. Note that the application requests can include requirements to the protection of the circuit.

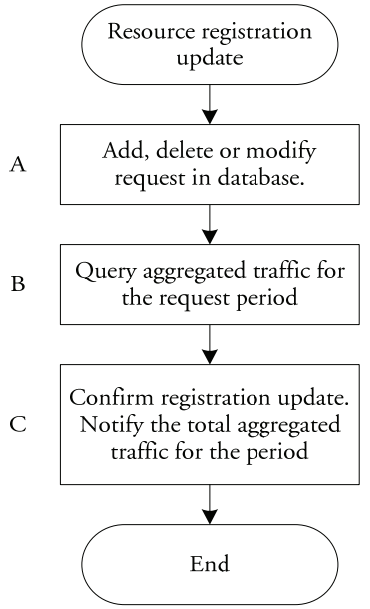


Fig. 6.6: Resource registration algorithm for registering and deletion of application requests

The execution of the algorithm is triggered by requests from the application through the API and the NSP. This is done whenever the user or application sends a new application request or wants to update or delete an existing one.

First, in A, the application request parameters are stored in the database, and an ID is added to the request. If an existing application request is modified, the call to the algorithm should include the ID or handle for the request to modify.

After registering the request, the total bandwidth demand and number of request for the start time of the application are queried as shown in Fig. 6.6 (B). If the bandwidth is expressed in both peak and average bandwidth, a simple aggregation scheme as defined in [98] could be used, where the peak and average bandwidths are simply added. If only the bandwidth is expressed as a single parameter this is simply added as well.

The objective of querying the aggregated traffic demand is to provide an indication to the NSP and the application whether the application demand will actually be satisfied. Hence, the aggregated traffic parameters are returned to the application for new or modified application requests (C). It is important to stress that the response only confirms that the request has been registered and thus not that the requested resources will eventually be available.

6.3.1.2. Look-ahead reservation concept

The actually resource allocation in the circuit layer is triggered from the look-ahead or resource triggering function. This function periodically looks ahead and determines if the actual reserved resources in the circuit layer are sufficient to satisfy the pending resource requests. If further resources are required, these are requested through the UNI and vice versa; if unused resources are allocated these are released. The resource triggering algorithm is shown in Fig. 6.7. The algorithm is executed periodically independent of the resource registration algorithm. The time period between executions is flexible and determined by the network operator.

In (A), the database comprising the application request is queried to identify the required resources in near future. More specific, the maximum requested resources within the *preAllocation* and the *preRelease* time periods are queried. These parameters, further discussed in section 6.3.2, partly determine the granularity and level of dynamics of the network. I.e., if the current time is given by T then the maximum aggregated requested resources in the intervals $[T;\ T+preAllocation]$ and $[T;\ T+preRelease]$ are queried from the database. In addition, the actual available resources in the circuit layer are queried from the network directly through the UNI or through the UNI proxy. These pieces of information are sufficient to determine the remaining flow of the algorithm.

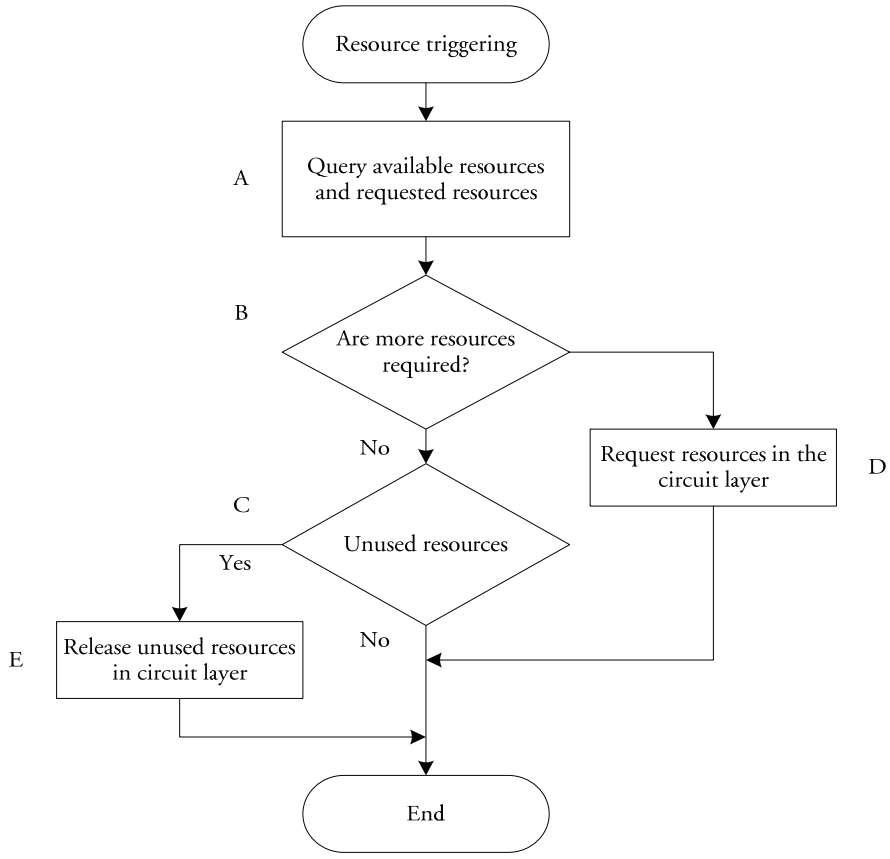


Fig. 6.7: Resource triggering. Based on the aggregated registrations resources are allocated from the circuit layer.

In Fig. 6.7 (B), the requested resources within the *preAllocation* period are compared with the actual resources. If more resources are required, these are requested from the circuit layer (D). Note that the algorithm-flow is independent whether the resource request to the circuit layer succeed or fail; if the resource request fails, then a new attempt to request the resources is done in the next periodic execution of the algorithm.

If no more resources are required, it is in (C) further investigated whether unused resources should be released. This is based on a comparison between the actual available resources and the requested resources within the *preRelease* period. If resources can be released these are released in the circuit layer (E).

Obviously, a cost model should be enforced to motivate the users of the network to release resources when possible; however, such a model is not within the main scope of this work.

6.3.1.3. Reservation notification algorithm

The resource registration algorithm registers the application requests and the resource triggering algorithm requests resources in the circuit layer accordingly.

The notification algorithm has two main objectives. It verifies whether the requested capacity is available within a certain period before the capacity should be used, and it notifies the NSP accordingly for each application request. The notification algorithm is shown in Fig. 6.8.

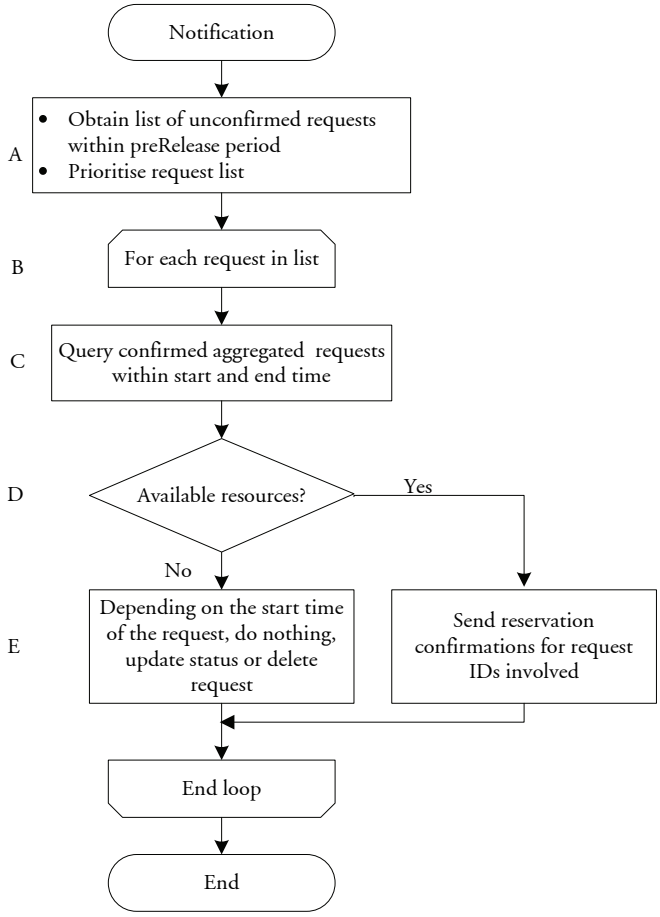


Fig. 6.8: Notification of NSP whether reservations have been satisfied or not. The algorithm is executed periodically after the resource triggering or by request.

A list of non confirmed application request within the *preRelease* time period is obtained from the database, and the list is sorted based on prioritisation (A). The simple prioritisation approach is based on first registered have highest priority. The *preRelease* parameter is used here, because if any requests are present within this period it is known that the resources will not be released within the *preRelease* period.

Then, each application request is treated separately with the highest priority processed first. In (C), the aggregated registered traffic for the existing requests are queried to determine if resources are available and free. It is noted that only the already confirmed requests are queried, which differs the query from the query in section 6.3.1.1.

If resources are available for the full duration of the request, the status of the request is set to *confirmed*, and a message is sent with the request ID and the status to the NSP (D).

If the resources for the full duration of the request are not available the following actions are taken (E).

- If the time to the start time of the application request is longer than the *preAllocation* parameter, no action is taken. This is a normal situation and it is expected that the resource triggering function will attempt to reserve the resources in the interval [T_{start}-*preAllocation;* T_{start}].
- If the time to the start time is less than the *preAllocation* parameter and more than the *deadline* parameter, the status is modified to *pending* and a notification message is issued to the NSP. This informs the client application that reservation of the resources have been attempted, but until now has resulted in failed setup. The situation is not fatal because the resource triggering algorithm will reattempt to reserve the resources on its next periodic execution.
- If the time to the start of the request is less than both the *preAllocation* and the *deadline* parameter, the request cannot be satisfied and the status is changed to deleted or postponed; in the latter case a new start time is inserted. An error notification is issued to the NSP.

As described, no hard guarantees are offered to the client application, but the probability of a successful resource allocation is highly dependent on the *preAllocation* and the *deadline* parameter. Whether to postpone a registration depends on whether this option is indicated in the application request. For a certain grid application where the network resources are reserved in combination with supercomputing resources it might not be possible to postpone the registration and it should be up to the client to request a new time slot.

6.3.2. Specifying level of dynamics

The three algorithms, "registration", "triggering" and "notification" presented in the previous section specify the level of dynamics in the network. Generally, setting the values of the *preRelease* and the *preAllocation* low increases the dynamics of the and vice versa; setting these values high makes the network more static.

The *preAllocation* parameter is specified indirectly by the applications as a predefined value is associated with the requested class of service. It could also be an option to allow the client to specify the *preAllocation* parameter without constraints; however, such approach makes it less possible for the network operator to control the level of dynamics in his network.

The *preRelease* parameter is the key parameter for controlling the dynamics, why it is controlled solely by the network administrator of the network domain providing the UNI. The *preRelease* parameter specifies how long in advance no requests are registered in order to release resources.

In addition, the periodicity of the algorithm executions also sets a lower bound on the dynamics of the network. If it is desired to have the dynamics in the minute scale, then the resource triggering algorithm should be executed at least once pr. minute. It is presently not clear how much processing and signalling overhead such approach provides, why modelling activities are initiated to illustrate this. The first bounds on the dynamics are modelled in section 6.4, based on the reconvergence of the routing protocols.

For multi domain networks, some extra considerations should be done. Each network operator is only responsible for controlling the dynamics of his own network domain; however, the dynamics of multi domain edge to edge connections depends on the dynamics of all the networks involved in the connection. It is thus preferred if the network domain provide dynamics parameters that are comparable for the multi domain connections. On the other hand, even with different levels of dynamic of the different network, the concept will still be operational, although the level of dynamics for multi domain connections mainly will be determined by the network with the lower level of dynamics, i.e., the static network or the network with highest values for the *preRelease* and *preAllocation* parameters.

A conservative operator who wants to implement dynamics in his network can use the proposed scheme setting the *preRelease* and the *preAllocation* parameters sufficiently high, which equals a network that is close to static. Then the dynamics can be increased in small steps, which is probably more appealing than implementation of complete dynamic infrastructure from day one.

6.3.3. Parameter description

Each application request comprises parameters to detail the traffic characteristics and class of service it requests. These parameters are start time, end time, the service definition parameters and the QoS specification. A list of the parameters to store in the database is provided in Table 6.2.

Table 6.2: Request parameters description for a packet request

Parameter	Type	Comments
Basic parameters:		
Request ID	Integer	Unique identification number for the request.
Destination Edge Address	IP	Address to identify the edge of the circuit switched network that the flow belongs to. This is not the end IP address, but the edge of the circuit switched network
Start time	Time	Identifies the start time of the connection request. A value of zero means setup immediately..
End time	Time	Identifies the end time of the connection request. For a value of zero the application SHOULD send a release message for the request ID explicitly.
Flow specification:		
Bandwidth	Integer	Specifies the bandwidth to be supported.
Delay	Integer	The maximum allowed delay of the connection.
Link protection	Byte	Specifies the amount of protection required in the circuit layer, e.g., if a disjoint path should be established.
Setup parameters		
Service Class	Integer	Defines priority for the service. The selection of the service class also selects within a set of *preAllocation* values.
DeadLine	Time	Time in advance of the start time the connection the available resources should be available.
Postpone	Bool	Indicates whether the reservation should be retried at a later time if resources cannot be reserved. If not, the registration is deleted if setup fails.
Status	Byte	Identifies whether the connection is confirmed or not, i.e., if the resources are available in the circuit layer and the application has been notified

The parameters are separated in three different sub groups. The basic parameters contains information of edge addresses for the circuits and information of start and end time. Then the flow specification sub groups include information of the traffic characteristics and minimum requirements of the application requests. Finally, the setup parameters specify service class (indirectly selecting *preAllocation* parameter), deadline value and whether the request can be postponed if the resources cannot be provided.

The parameters, *request ID* and *status*, are provided by the algorithms in section 6.3.1, and they give feedback to the requestor. Obviously, it is not feasible to allow an application not to define an end time. However, a proper cost model would provide motivation to the application to send an explicit release message to the adaptation function after using the connection.

6.3.4. Architecture

The flow of information within the advance reservation block is illustrated in Fig. 6.9, where the horizontal dashed lines marks the functionality of the advance reservation block.

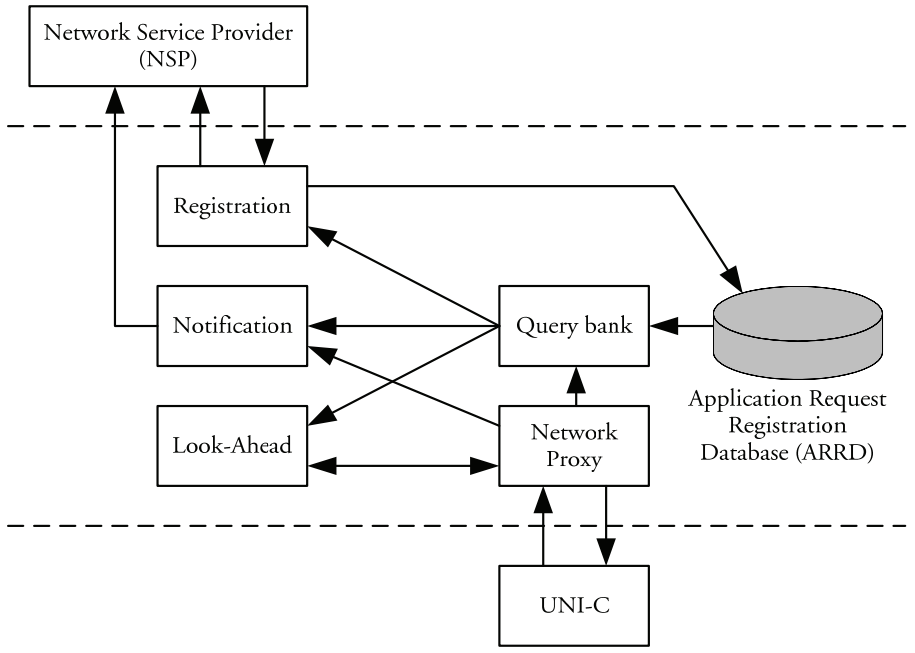


Fig. 6.9: Block diagram for advance reservation block. The arrows indicate main data flows.

The main sub blocks in the advance reservation block are the blocks corresponding to the three main algorithms, the query bank and the network

proxy. In addition a database of a certain size is required to store the present reservation requests.

6.3.4.1. Registration Entity

Application requests are transferred from the web services based API service provider to the registration entity.

The registration block labels each request with a *Request ID* and stores it in the Application Request Registration Database (ARRD) with the parameters shown in Table 6.2. The status field is set to *registered*.

The ARRD stores the registration requests independently of the current available resources in the circuit layer and confirms the registration. Upon receiving the registration confirmation the registration entity request the total aggregated registered bandwidth between the start and end time of the application request in addition to the currently available bandwidth. This is requested from the query bank that queries the database and the network proxy, which is aware of the current circuit resources.

6.3.4.2. Look-Ahead Entity

The triggering of new circuits and release of circuits are done through the Look-Ahead Entity (LAE), as described previously. The LAE communicates with the Network Proxy and the Query Bank.

The connection between the LAE and the ARRD is done only through the Query Bank. Furthermore, the LAE uses the Query Bank to query the available resources through the Network Proxy. The only information flowing directly between the LAE and the Network Proxy are the request for releases and allocation of new resources in the circuit layer.

Only the LEA can request establishment or release of connections, and it is executed periodically.

6.3.4.3. Notification Entity

The application or the client user should know when the connection is finally confirmed or rejected. This is done through the Notification Entity (NE) which is logically connected to the network service provider (NSP), the Query Bank and the Network Proxy.

In the algorithm in Fig. 6.8, a notification is generated and sent to the network service provider interface. This message should contain the following pieces of information:

- Request ID.
- Status of the application request:

  - *Confirmed*: Indicates that the request has been accepted and the required resources have been allocated. The actual confirmed traffic parameters are provided in the value field.
  - *Registered*: The application request has been registered, but no attempts for registering the resources have yet been done. The amount of other traffic and the available bandwidth is provided in the value field.
  - *Pending*: The resource allocation has been initiated but it is not complete. This is due to the delay in the reservation signalling or because reservation attempts have failed. The number of attempts is provided in the value field.
  - *Postponed*: The original reservation request has failed but it will be retried in near future. The new start time is provided in the value field.
  - *Failed*: The reservation has failed.
- Value field: Flexible field used to provide further explanations for the status field.

In this section it is only considered how the notification message is communicated to the NSP. It is thus an open question how the NSP relays this information to the application. The most obvious solution is that the application is responsible to query the NSP for the information. This stems from the fact that it is not suitable to keep a web services session open to wait for a status update several days ahead.

6.3.4.4. Query Bank

The Query Bank is located between the ARRD and the registration, the look-ahead and the notification entities. Basically, the Query Bank comprises a number of queries on the ARRD and the Network Proxy.

The registration block uses the Query Bank to query the total aggregated bandwidth that is registered in the database in combination with the actual bandwidth in the circuit layer. The look-ahead entity uses the Query Bank to obtain the total aggregated traffic in comparison with available bandwidth analogous to the registration entity. However, instead of querying for the start time of the application requests the start time minus the *preAllocation* time is used.

When called from the NE, the Query Bank should also be able to provide a full list of application request within a given time period.

6.3.4.5. Application Registration Request Database

The Application Registration Request Database (ARRD) is a storage base containing the necessary registrations and including the required parame-

ters for *preallocation* and *deadline*. The ARRD can be accessed directly or through the Query Bank.

6.3.4.6. Network Proxy

The main objective of the network proxy is to relay the resource request and the resource release from the look-ahead entity. Secondly, the proxy keeps track of the actually reserved and available resources, a piece of information used by the aggregation and query functions.

6.3.5. Examples

In this subsection a couple of examples will be provided to illustrate the flow of the suggested resource triggering approach

6.3.5.1. Single scheduled reservation

A videoconferencing session (50 Mbit/s) is planned between 14.00 and 17.00 a specific day. The *deadline* parameter is set to 1 hour and the *preAllocation* parameter to 2 hours. The reservation request is registered at 7.00.

Consider no other reservations are registered within the time interval from 14.00 to 17.00, and the available bandwidth is 100 Mbit/s. Further consider that the *preRelease* parameter is set to 12 hours by the network operator.

In this situation the reservation is guaranteed already after registration as the available resources are present and no release will take place because the *preRelease* parameter is higher than the time in advance the registration is made. The registration is confirmed.

Alternatively, consider the situation where only 20 Mbit/s is available in the transport layer. In this case the situation is as follows:

At 12.00 (2 hours before the session) the look-ahead function discovers that the requested resources are larger than the available. Hence, the transport layer is requested to provide extra capacity in granularities depending on the transport technology. Reservation requests are transmitted from the look-ahead function through the Network Proxy and the UNI to the transport network. In case the requested resources are allocated, the notification function generates a confirmation message and tags the videoconferencing as confirmed. In its next periodic execution the look-ahead function detects that available resources are present and no further actions are taken.

In case the network responds that no or too little resources can be allocated the notification function will tag the video conferencing request as “pending”, and the look-ahead function will retry the resource request until it receives the resources or until 13.00 (Start time minus *deadline* parameter). In

this case the notification function will invalidate the request and mark it "failed".

6.3.5.2. Local resource control

If no local control is included, the above scheme is complete analogous to "scheduled on demand" reservation, because no guarantees are provided before the reservation is actually made in the transport layer. This is inherent for the UNI interfacing.

The main difference is the local control in the adaptation function, which is illustrated by the following example:

Reconsider the video conferencing request from section 6.3.5.1. Add to the video conferencing request, a video streaming request with start and end times of 13.00 and 15.00, respectively, is registered at 9.00 with *preAllocation* parameter of 2 hours and deadline parameter of one hour. The parameters are shown in Table 6.3.

Table 6.3: Parameters for two application requests.

	Video streaming	Video conferencing
Registration time	9.00	7.00
Start time	13.00	14.00
End time	15.00	17.00
PreAllocation Parameter	2 hours (11.00)	2 hours (12.00)
Deadline Parameter	1 hour (12.00)	1 hour (13.00)
Bandwidth	50 Mbit/s	50 Mbit/s

Consider the situation where 80 Mbit/s is available in the circuit layer, which is insufficient to satisfy both applications. When the video conferencing session is registered this is confirmed in alignment with the example in section 6.3.5.1. This is, however, not the case with the video streaming request where only 30 Mbit/s is free, which is 20 Mbit/s less than required. Hence, at 11.00 the look-ahead function discovers that bandwidth should be allocated for the video streaming application. Whether or not the look-ahead function succeed in allocating bandwidth, the Video Streaming request is confirmed or marked as failed.

Hence, the local control ensures the possibility of providing priority to application requests based on first registered or any other mechanism.

6.4. Modelling level of dynamics

In the previous section, a resource allocation concept was proposed. The concept is compliant with a distributed network based on UNIs and signalling with RSVP or CR-LDP. The level of dynamics of each network domain is primarily determined by the values of two to three handles, namely the *preAllocation* and *preRelease* in addition to the periodicity of the execution of the proposed algorithms.

However, what is the absolute highest level of dynamics that should potentially be expected from a network of a certain size? I.e., which are the shortest time scales for which it is beneficial to allocate resource demands?

This section addresses these questions by observing the effect of the reconvergence time of a traffic engineered network. If the network is in a reconvergence state, then it is not useful to establish connections before it has converged. Therefore, this section briefly describes the *Open Shortest Path First TE* (OSPF-TE) signalling protocol with traffic engineering extensions and its default parameters. Then the convergence times in a network comparable in size to the GÈANT network is evaluated to indicate general reconvergence times. The section in mainly based on [Publ. 16], [Publ. 14], [Publ. 18] and [Publ. 19].

6.4.1. OSPF for traffic engineering

Open Shortest Path First (OSPF) is an Interior Gateway Protocol (IGP), originally developed by the IETF for IP networks. Current version, OSPF version 2, is described in [99]. Developed for MPLS/GMPLS based networks, OSPF is extended to support traffic engineering known as OSPF-TE [100]. A new type of *Link State Advertisements* (LSA), Opague LSA, is deployed to describe the network topology and bandwidth information. In addition to standard OSPF, the Opague LSA carries information of the total and available bandwidth for a given link. LSAs are thus distributed whenever the reserved bandwidth for a link is changed with more than usually 10% of the total capacity. After generating an LSA, the router floods the LSA to all its adjacencies. Upon receiving an LSA, a router updates its own Link State Database (LSDB) according to the recent change, duplicates that LSA and re-advertises the LSA copies to all neighbours. The reliable flooding procedures allow every router to maintain a consistent snapshot of network topology and link information.

6.4.1.1. OSPF Timer Configuration

The functionality of OSPF is heavily dependent on how its timers are configured. These times specifies when a new LSA update should be distributed and how incoming flooding messages should be handled. For standard OSPF the default values are described in [101] and [102], and the most important values are provided in Table 6.4

Table 6.4: Most important default values of the OSPF timers in simulation

Parameter	Value	Description
Hello Interval	10	The time interval between hello packets sent from each interface.
Router Dead Interval	40	The time threshold beyond which an interface not responding to hello packets is declared inoperable.
Transmission Delay	1	The time it takes to transmit an LSA to neighbours on the interface.
Retransmission Interval	5	After this time, if no acknowledgement is received, the interface retransmits the last packet transmitted.

The routers generate LSA messages whenever the LSA contents change. Thus, the traffic engineering LSAs are generated whenever the link state information is modified. In OSPF-TE [100], the flooding of updates is not required immediately following every change. Three mechanisms exist: the immediate update mechanism, the threshold-based update mechanism [103] and the period-based update mechanism [104]. In the immediate case, any modifications in the LSAs are advertised immediately. In the threshold-based case, a threshold (for example, the variance of bandwidth change) is set and flooding is triggered if this threshold value is reached. In the period-based case, a timer is set and all routers advertise their LSAs periodically. The first mechanism causes heavy control overheads due to the frequent changes of link states. This problem is solved by using the latter two mechanisms. However, they disseminate information with delay and will cause the inaccuracy of link state.

6.4.1.2. OSPF Network Convergence Time

The establishment of label switched paths depends heavily, whether the edge router to setup the route has a consistent view of the network. In other words; the LSP setup cannot be initiated before the network is completely stable and the routing protocols have converged. In the MPLS/GMPLS-based network, the convergence time at the ingress LERs is particularly im-

portant. Setting up paths before the network has converged inherently leads to failed path, looping paths and over-utilisation of links causing delay and packet loss. Such deteriorated performance of the transport network cannot be tolerated by delay sensitive applications like, e.g., video conferencing.

The route convergence is composed of the LSAs exchanging time, the link state databases updating time and the route calculation time. The following lists some influencing parameters that must be taken into account when calculating the OSPF convergence time.

- **Network topology and connectivity**. The flooding load of OSPF protocol increases, when the size of network becomes large.
- **Traffic demand patterns**. The more and shorter traffic requests, the more flooding of routing information must be expected.
- **Current traffic load**. The LSA messages can be dropped due to the network congestion if the links are already over-utilised. The loss of LSA information will further delay convergence.

6.4.2. Modelling convergence with OPNET

One of the factors for the convergence time in OSPF-TE is the size and density of a network. Other factors are the configuration of the timers and the load in the network. In the following the convergence time for a small network and a network domain of approximately 25 core nodes are evaluated. The latter topology is based on the main links and routers in the GÉANT2 network [89].

The implementation of the modelling is done through the protocol and network modelling tool OPNET [105]. A threshold for the LSA distribution is implemented, which specifies that the routers should flood LSA updates if the bandwidth utilisation of a link is changed with more than 10%. In the following simulation scenarios, the links connecting routers are DS1 links (1.544 Mbps) and the threshold value is 158 Kbps. When the reserved bandwidth value is larger than the threshold value, TE LSA messages are advertised immediately.

6.4.3. Verification in small networks

The first scenarios for evaluating the convergence time are for small networks of five nodes.

In the first scenario a full mesh is used, why it is expected that the convergence time will be short as all routers are connected directly. An LSP is established between two edge nodes at T+150s, where T denotes the start of the simulation. Because the LSP allocates more than 10% of the link capac-

ity flooding of LSAs are initiated, and the convergence activity lasts for 2.3 ms.

In the second small verification scenario, the number of links is reduced and the LSP setup is repeated. Now, convergence activity is observed for 3.5 ms.

Hence, for this very simple network a slightly increase in the convergence time is observed by reducing the density of the network.

6.4.4. Convergence in large networks

The OSPF convergence time is simulated in a large network topology based on the GÉANT2 network [89] illustrated in Fig. 6.10.

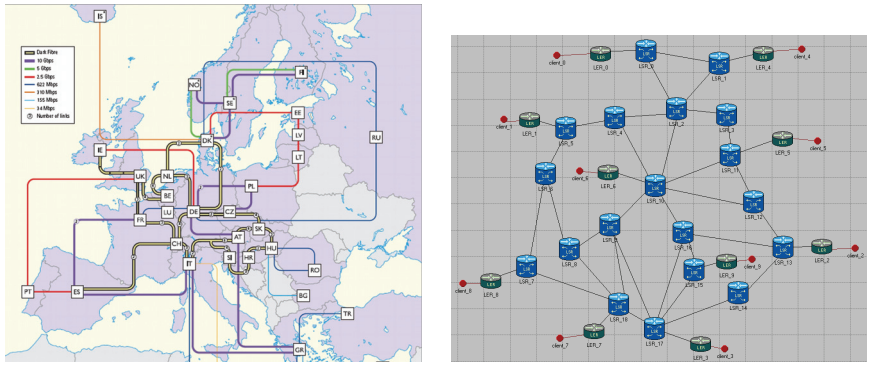


Fig. 6.10: GÉANT2 network topology [89] and the derived implementation of the topology in OPNET. The green LERs are added afterwards to enable LSPs from each LER to any of the other LERs.

In this case, LSPs are one directional and configured between the edge-routers. There are in total 90 LSPs defined between the LERs. The results are shown in Fig. 6.11, where the average and maximum values of network convergence time are calculated for each LER. The convergence time is approximately 10 seconds, why this is the absolute minimum time intervals to consider for a dynamic infrastructure.

It was chosen to use a topology close to the GÉANT2 topology as this obviously reflects the topology used for European high demanding research applications.

It is seen from the results in Fig. 6.11 that the reconvergence time depends heavily on the size of the network in terms of node count. E.g., the LER0 results are the maximum and the average reconvergence time for all the LSPs originating from LER0 to all of the other LERs subsequently. Each reservation equals half the bandwidth of the link.

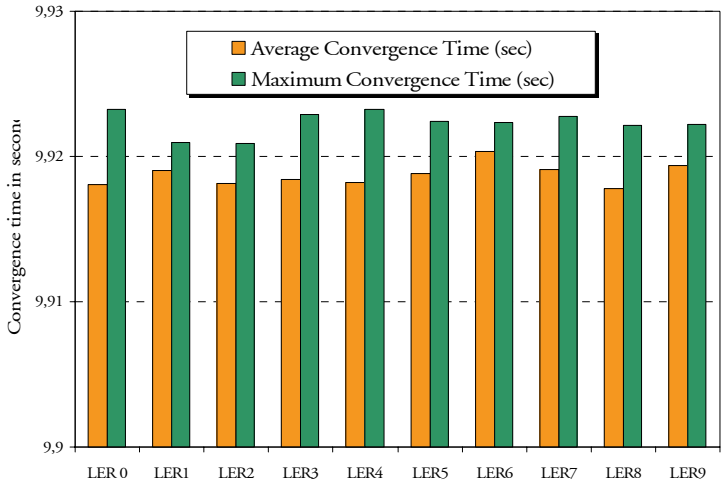


Fig. 6.11: OSPF network convergence time for large networks

As expected, the convergence time only depends slightly on which LSP is modified. This is because the link state information needs to flood the complete network in all cases. Hence, it is obvious that the convergence depends on the network size. For a network of approximately 25 nodes corresponding in size and topology to the European Research Network GÉANT2 the convergence time under normal operation equals 10 seconds.

6.5. Summary and future perspectives

The capacity and flexibility of the packet switched IP networks are directly dependent on the dynamics of the underlying circuit switched transport networks. The performance of high network demanding applications is improved if these can request resources in the network; not only in the packet layer but also in the underlying circuit layer.

In this chapter, application triggered dynamic bandwidth allocation in a circuit switched transport network has been considered. It is assumed that the heterogeneous multi-domain transport network includes an interoperable control plane with standardised user-network interfaces.

The ongoing European IST project MUPBED addresses both the horizontal integration of several multi-vendor and multi-administrated network domains, and the vertical integration of the applications and the network control layer. The work described in this chapter is closely connected to the MUPBED project.

The communication between the applications and the network control plane is defined by an adaptation function with a web services based API at the application side and a UNI at the network side. This function includes the advance resources allocation function, which accepts requests from the applications and reserve resources accordingly.

Three main sub functions were proposed to implement the advance reservation function. One receives and stores the application requests, one triggers resources in the network layer through the UNI based on the aggregated reserved resources and, finally, one notifies the application layer, whether the request is satisfied. In addition, the proposed concept supports local priority, which can be used to give priority to connection requests, e.g., first come first served etc.

The specified model provides for the network operator a few and well defined parameters for controlling the level of dynamics in the network. By controlling these values, he can specify whether the dynamicity should be in the minute, hourly or daily scale.

Finally, the highest granularity of the dynamicity, i.e., how short time scales should a circuit be allocated for, is modelled by observing the reconvergence times for OSPF-TE routed networks. For a research infra structure comparable to the European Research Network GÉANT, a minimum reconvergence time of 10-20 seconds are measured, why a time granularity below this would lead to failed LSP establishments.

Because the MUPBED project is still ongoing, the resources triggering scheme has not yet been finally implemented for large scale networks. However, the guidelines in this chapter provide a clear framework for the implementation, and the first light implementations are used to let a few selected applications communicate the resources requests to a multi domain network.

While the current work is focusing on the network support for high demanding research applications, it is believed that the results can also be used to implement dynamics in commercial and residential networks, as most of the current research applications like video-conferencing will be common residential applications of the future. Hence, the schemes are used for improving the utilisation of current network, while providing application specific QoS. Improved utilisation and dynamics of transport network without human interaction can significantly reduce the cost of operating the networks.

7. Conclusion

After the financial bubble for the IT related industries in the beginning of this century, the road forward has been to use the already deployed capacity as cost-efficiently as possible, while improving the user perceived quality.

In this thesis, different topics for controlling optical infrastructures have been investigated, and improvements have been suggested. Either potentially more cost efficient solutions or better utilisation of the existing resources are proposed.

Again it is important to understand the performance of the network as it appears to the application user, which might easily differ from figures like bit error rate and other more transport related parameters.

The challenges operating optical infrastructures in the core and access networks are quite different. In the core network the cost of the components are shared by many customers and the main design parameters are scalability, capacity and reliability. In the access network, however, the main objective is to use components that are suitable for high level of integration with potential huge savings.

Optical packet switching has been suggested as a future technology for the core network combining capacity and scalability. In relation to this, optical signal regeneration has been addressed, where a signal from an optical packet switch is regenerated before further transmission, i.e., the signal is properly amplified and noise and timing jitter is reduced. The optical packet switch considered in relation to the European IST project DAVID results in a packet stream at the regenerator with packet-to-packet power level variations up to 5 dB. This is unacceptable for the optical regenerator, and the packet-to-packet power level should be equalised before the regenerator.

To solve this, an optoelectronic power level equalisation device was developed. Two approaches for the power control were assessed; controlling the gain of an input arm of a Mach-Zender Interferometer used for the regeneration or controlling the gain of an SOA operating in XGM regime prior to the regenerator. The second option was chosen as it was independent on the optical signal to noise ratio of the incoming signal.

The power equalisation was developed with digital electronics, as this is the only approach to efficiently distinguish between a guard band and long sequences of zeroes in the signal. The control electronics determines the power level based on the fixed formatted preamble of the DAVID packets. The determined power level is then assumed for the duration of the packet, a feature not possible with the proposed solutions in the literature. Furthermore, compensation against a non-linear transfer function of the laser to control was implemented in the digital control electronics.

The combination of the signal regenerator and the control electronics was able to efficiently equalise optical packet-to-packet power variations up to 9 dB. In addition, the adaptation to a new power level was done within 10 ns, which is perfectly within the guard band between two packets of approximately 50 ns. The performance measurements were assessed with a bit rate of 10 Gbit/s, however, the control electronics is completely independent on the bit rate why it is argued that the design can be used for power equalisation of 40 Gbit/s and above.

Standard label swapping in optical packet switched networks requires header modification in the core nodes. However, schemes for such processing are far from mature, and it is beneficial if header modification in optical packet switched network is avoided.

A solution to this problem is the Key Identification Scheme, which is a novel scheme based in the Chinese Remainder Theorem. Instead of modifying the header for each traversed node, all the forwarding information is encoded in a label, and each node extracts the correct addressing information using a function on the label and a node-specific and unique key, which is distributed when the network was initialised.

The approach, however, results in labels larger than standard MPLS labels, why the scalability of the scheme has been addressed by simulations. These showed that a label size of up to 6 bytes should be expected if the scheme is deployed within a network with up to 50 optical core nodes. For even larger networks the scalability was analysed and if the route length is restricted to less than eight nodes, a network of up to 200 nodes is supported with a label length of 8 bytes.

The required function in the electronic header processing of the optical packet switch is a modulo function on the label and the node specific key. It is demonstrated how such function can be computed is standard digital electronics within a few clock cycles, well within the duration of a DAVID packet through the optical packet switch.

Through the KIS, it is thus demonstrated that header modification in optical packet switched networks can be avoided with negligible impact on the

scalability towards network size. This is believed to bring optical packet switching one step further towards commercial realisation.

Considering the access network, it seems evident that the user is not willing to pay a lot extra for new high bandwidth services. This demands very cost efficient solutions in this part of the network.

Therefore, candidate components for future Fibre to the Home (FTTH) systems were evaluated with special focus on the lasers, the receivers and the front-end electronics. The evaluations were based on a case study comprising a four channel 10 Gbit/s point-to-point system.

For both the 13xx nm and the 15xx nm wavelength region, directly modulated lasers were assessed, because these are simpler than externally modulated lasers and thus potentially more suitable for high level integration.

The laser performance was evaluated in the case study with focus on the power level and the dispersion characteristics. Both lasers were suitable for 10 Gbit/s transmissions in a 10 km FTTH system, and the dispersion effect for the 1550 nm laser was only indicated in range of 25 km. using standard single mode fibre.

Furthermore, it was measured that the wavelength drift with temperature variations was about 7 nm for a 70°C variation. This is well within the pass band of the CWDM devices, which was recommended for the separation of the wavelength channels.

It was not possible to conclude whether one should use 13xx or 15xx lasers for future FTTH systems, and this will depend on the monolithically integration properties of the laser material for the two regions.

The results document the viability of directly modulated lasers for fibre transmission in the access loop, which is expected to improve the yield for high level integration of the optical components and thus reduce the deployment costs for FTTH systems to the residential users.

Today, the circuit switched optical infrastructure and the packet switched data layer and the applications are working independently. The data applications can utilise the allocated circuit switched resources, but they cannot request new resources or adjustment of existing resources. By integrating the application with the circuit control layer is believed to improve the utilisation of the physical available resources.

This has been addressed as part of the activities in the European IST project MUPBED, which considers the horizontal integration of multi-domain network and vertical integration of application with the network control plane.

The vertical integration is addressed in this thesis, focusing on highly demanding research applications. The applications communicate with an adaptation function, which translates and forwards the request to the network control plane, and if everything proceeds as planned the network allocates circuit switched resources for the applications. In this way a dynamic bandwidth allocation is established in the transport network, which increases the utilisation of the network, while ensuring guaranteed bandwidth for certain applications.

The resource triggering is based on three algorithms which register resources, reserves resources based on the aggregated application requests and notification of the application layer with the result of their resources requests. A few key parameters are provided for the network operator to easily control the level of dynamics of his network.

The highest level of dynamics possible for a typical research network was evaluated through modelling based on the reconvergence times of the OSPF-TE routing protocol. This showed that after a significant change in the network at least 10-20 seconds was required, before new establishments of MPLS LSPs should be initiated. This indicates a practical granularity in the minute-scale.

One important thing, which has been the driver for the many different topics in the thesis, is cost and integration. After the bubble collapsed, the focus has shifted from capacity to cost; especially the studies in the access network support this.

This also slightly moves the focus from all-optical to electro-optical solutions. The thesis has addressed two major challenges in optical packet switched network that can be solved using simple electronic controlling schemes. If it is believed that optical packet switching again will be on the agenda, such solutions should be used to speed up the transition from the lab to the field, and especially the Key Identification Scheme addresses a very complex optical problem with a simple electronic approach.

It was asked in the introduction, which triggers should be expected for the future of the data and telecom industries. In the access network, the users are only willing to pay marginally extra for new services. Here, extreme cost-efficiency can provide a little revenue which can be used for the necessary and related upgrades in the core network. For research and business users the stochastic delay variations in the packet switched network are not suitable for critical interactive services, which require high bandwidth in combination with low delay. It is believed that providing prioritised premium services, these users are willing to pay significantly more than for a best effort service. The contribution in this thesis related to the vertical in-

tegration indicates how such services can coexist with the best effort services.

Thus, the results in this thesis, which have been disseminated in international journals and conferences, demonstrate viable solutions for cost-efficient fibre access, separation and high utilisation of the optical infrastructure and electronic solutions for two critical issues of optical packet switching.

8. References

[1] G. P. Agrawal, *"Fiber Optic Communication Systems"*, Wiley-Interscience, 2nd Edition, 1997.

[2] N. McKeown, "Scalability of IP routers", *in proc. of Optical Communication Conference (OFC) 2001*, (invited), Anaheim, USA, 2001.

[3] L. Y. Lin, E. L. Goldstein and R. W. Tkach, "Free-Spaced Micromachined Optical Switches with Sub millisecond Swicthing Time for Large Scale Optical Crossconnects", *IEEE Photonics Technology Letters*, vol. 10(4), p. 525-527, April 1998.

[4] Official DAVID website. "Data and Voice Integration over DWDM". [online]. Available: http://david.com.dtu.dk

[5] MUPBED website, [online] http://www.ist-mupbed.org.

[6] D. Bruce, "MPLS Technology and Applications", *Morgan Kaufmann*, 2000

[7] U. Black, "MPLS and Label Switching Networks", *Prentice Hall*, 2001

[8] E. Rosen, "Multi Protocol Label Switching Architecture", *IETF RFC 3031*, January 2001

[9] C. Huitema, "Routing in the Internet", *Prentice Hall*, 1995.

[10] B. Jamoussi, Ed., L. Andersson, R. Callon, R. Dantu, L. Wu, P. Doolan, T. Worster, N. Feldman, A. Fredette, M. Girish, E. Gray, J. Heinanen, T. Kilty, A. Malis: "Constraint-Based LSP Setup using LDP.", *IETF RFC 3212*. January 2002.

[11] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow. "RSVP-TE: Extensions to RSVP for LSP Tunnels." *IETF RFC 3209*, December 2001.

[12] P. Ashwood-Smith et al: "Generalised Multi-Protocol Label Switching (GMPLS) Signaling. Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", *IETF RFC 3472*, January 2003.

[13] M. Berger, H. Christiansen, B. Mortensen and R. Jociles-Ferrer. "Hierarchical Electro-optical Packet Network Architecture", IST 2003, pp. 311-315, Isfahan, Iran (2003).

[14] ITU-T: "Architecture for the automatically switched optical network (ASON)", *ITU-T Recommendation G.8080/Y.1304*, November 2001.

[15] N. Larkin, "ASON and GMPLS – The battle of the optical control plane", *Data Connection,* [online] http://www.dataconnection.com, August 2002.

[16] OIF: "User Network Interface (UNI) 1.09 Signaling Specification, Release 2", *Optical Internet Forum*, February 2004.

[17] C. Guillemot, M. Renaud, P. Gambini, C. Janz, I. Andonovic, R. Bauknecht, B. Bostica, M. Burzio, F. Callegati, M. Casoni, D. Chiaroni, F. Clerot, S. L. Danielsen, F. Dorgeuille, A. Dupas, A. Franzen, P. B. Bukhave, D. K. Hunter, A. Kloch, R. Krähenbühl, B. Lavigne, A. Le Corre, C. Raffaelli, M. Schilling, J.-C. Simon, L. Zucchelli, "Transparent optical packet switching: The European ACTS KEOPS project approach". *IEEE Journal of Lightwave Technology*, vol 16, pp. 2117-2134, Dec 1998

[18] T. Fjelde: "Traffic analysis and signal processing in optical packet switched networks". *PhD Thesis*, Research Center COM, Technical University of Denmark, October 2001

[19] D. Hunter, M. Nizam, M. Chia, I. Andonovic, K. Guild, A. Tzanakaki, M. O'Mahony, J. Bainbridge, M. Stephens, R. Penty and I. White, "WASPNET: A wavelength switched packet network", *IEEE Communications Magazine*, pp. 120-129, March 1999.

[20] A. Carena, M. D. Vaughn, R. Gaudino, M. Shell and D. J. Blumenthal, "OPERA, An Optical Packet Experimental Routing Architecture with Label Swapping Capability", *IEEE Journal of Lightwave Technology*, vol. 16(12), pp. 2135-2145, December 1998.

[21] D. Wonglumson, I. M. White, S. M. Gemelos, K. Shrikhande and L. G. Kazovsky, "HORNET – A Packet-Switched WDM Network: Optical Packet Transmission and Recovery", *IEEE Photonics Technology Letters*, vol. 11(12), December 1999.

[22] B. Hajek and T. Weller, "Scheduling nonuniform traffic in packet-switching system with small propagation delay", *IEEE/ACM Transaction on Networking*, vol 5, pp 813-823, December 1997.

[23] M. S. Berger, V. B. Iversen, B. B. Mortensen, "Analytical performance evaluation of optical packet network interface", *Conference on Optical Internet Networking (COIN) 2003*, Melbourne, Australia, 2003

[24] B. B. Mortensen, M. S. Berger, "Estimating timeout parameters for packet aggregation", *Conference on Optical Internet Networking (COIN) 2003*, Melbourne, Australia, 2003

[25] M. Shreedhar and G. Varghese, “Efficient Fair Queuing using Deficit Round Robin”, IEEE/ACM Transaction on Networking, vol. 4(3), pp. 375-385, June 1996.

[26] L. Dittmann, H. Christiansen, M. Berger, “Hierarchical MPLS – An approach for efficient resource administration in multitechnology networks”, *NOC 2001,* Ipswich, England, 2001.

[27] S. L. Danielsen, B. Mikkelsen, C. Jørgensen, T. Durhuus, K. Stubkjær, ”WDM packet switch architectures and analysis of the influence of tunable wavelength converters”, *IEEE Journal of Lightwave Technology*, 15(2), pp 219-226, February 1997.

[28] S. L. Danielsen, C. Jørgensen, B. Mikkelsen, K. Stubkjær, ”Optical packet switched network layer without optical buffers”, *IEEE Photonics Technology Letters*, 10(6), pp. 896-898, June 1998.

[29] T. Durhuus, *“Semiconductor Optical Amplifiers: Amplification and Signal Processing”*, Ph.D. Thesis, Department of Electromagnetic Systems, Technical University of Denmark.

[30] D. Keller, P. Aribaud, D. Auvinet, M. DiMaggio, F. Dorgeuille, C. Drocourt, D. Legat, V. Dhalluin, F. Pommereau, C. Porcheron, M. Prunaire, P. Piatti, J. C. Remy, M. Renaud, N. Sahri, S. Silvestre, S. Squedin, A. Göth, J. Scherb, “Hybridization of SOA’s on Si platform for routing applications”, *Indium Phosphide and Related Materials Conference (IPRM) 2000,* pp. 463-466*,* Marcoussis, France.

[31] B. Lavigne, “Cascade of 100 optical 3R regenerators at 40 Gbit/s on al-active Mach Zender Interferometers”, proceedings We.F.2.6, *European Conference on Optical Communication (ECOC’01)*, Amsterdam, Netherlands.

[32] N. Le Sauze, D. Chiaroni, M. Nord, M. S. Berger, H. Christiansen, J. Fernandez-Palacios, Jesus Felipe Lobo, D. Careglio, J. Solé-Pareta, S. Spadero, A. Rafel, A. Hill, S. Sygletos, H. Skoufis. A. Stavdas, H. Lønsethagen, T. Olsen, F. Callegati, F. Neri, A. Bianco, G. Galante and M. Mellia, ”Network concepts validation and banchmarking”, *DAVID Deliverable D101 (Public)*, December 2003.

[33] B. Lavigne, “All-optical 3R regeneration based on semiconductor technology”, TuA1 (Invited), *Photonics in Switching (PIS’2002)*, Korea, 2002.

[34] S. Fischer, M. Dülk, E. Gamper, W. Vogt, E. Gini, H. Melchior, W. Hunziker, D. Nesset and A. D. Ellis, “Optical 3R regenerator for 40 Gbit/s networks”, *IEEE Electronics Letters,* vol. 35(23), November 1999.

[35] B. Lavigne, P. Guerber, P. Brindel, E. Balmefrezol and B. Dagens, ”Cascade of 100 optical 3R regenerators at 40 Gbit/s. Based on all-

active Mach Zender interferometers", *Proceedings of European Conference on Optical Communication 2001 (ECOC'01)*, Amsterdam, Holland, 2001.

[36] T. Otani, M. Suzuki and S. Yamamoto, "40 Gbit/s Optical 3R Regenerator for All-Optical Networks", *Proceedings of 27th European Conference on Optical Communication 2001 (ECOC'01)*, Amsterdam, Holland, 2001.

[37] G. Raybon, "Optical 3R Regeneration in 40 Gbit/s Pseudo-linear Transmission Systems", *Technical Digest, Optical Fiber Communication Conference 2003 (OFC 2003),* Atlanta, USA.

[38] S. Watanabe, F. Futami, R. Okabe, Y. Takita, S. Ferber, R. Ludwig, C. Schubert, C. Schmidt and H. G. Weber, "160 Gbit/s Optical 3R-Regenerator in a Fiber Transmission Experiment", *Technical Digest, Optical Fiber Communication Conference 2003 (OFC 2003),* Atlanta, USA.

[39] S. Nielsen, J. C. Yen, N. K. Srivastava, J. E. Rogers, M. G. Case, R. Thiagarajah, "A Fully Integrated 43.2 Gbit/s Clock and Date Recovery and 1:4 Demux IC in InP HBT Technology", *IEEE Journal of Solid-State Circuits*, vol. 28(12), December 2003

[40] H. Kamisuna, T. Shibata, K. Kurishima, M. Ida, "A 43-Gbit/s Clock and Data Recovery OEIC Using InP/InGaAs/HBTs", *Technical Digest of Optical Fiber Communication Conference 2003 (OFC 2003)*, Atlanta, USA.

[41] B. Sartorius, C. Bornholdt, O. Brox, H. J. Ehrke, D. Hoffmann, R. Ludwig and M. Möhrle, "All-optical clock recovery module based on self-pulsating DFB laser", *IEEE Electronics Letters*, vol. 34(17), August 1998.

[42] C. Bornholdt, J. Slovak and B. Sartorius, "Semiconductor-based all-optical 3R regenerator demonstrated at 40 Gbit/s", *IEEE Electronics Letters*, vol. 40(3), February 2004.

[43] B. Lavigne, P. Guerber, D. Chiaroni, C. Janz, A. Jourdan, B. Sartorius, C. Bornholdt and M. Möhrle, "Test at 10 Gbit/s of an optical 3R regenerator using an integrated all-optical clock recovery", *Proceedings of European Conference on Optical Communication 1999 (ECOC'99)*, Nice, France.

[44] D. Chiaroni, B. Lavigne, A. Dupas, P. Guerber, A. Jourdan, F. Devaux, C. Bornholdt, S. Bauer, B. Sartorius and M. Möhrle, "All-optical clock recovery from 10 Gbit/s asynchronous data packets", *Proceedings of European Conference on Optical Communication 2000 (ECOC 2000)*, Munich, Germany.

[45] D. Chiaroni, "Novel All-optical Multifunctional Regenerative Interface for WDM packet switching systems", *Proceeding of European Conference on Optical Communication 1996 (ECOC'96)*, Oslo, Norway.

[46] D. Chiaroni, N. Le Sauze, T. Zami, J.-Y. Emery, "Semiconductor optical amplifiers: A key technology to control the packet power variation", *Proceedings of 27th European Conference on Optical Communication 2001 (ECOC'01)*, Amsterdam, Holland.

[47] S. L. Danielsen, P. B. Hansen, K. E. Stubkjaer, M. Schilling, K. Wünstel, W. Idler, P. Doussiere and F. Pommerau, "All Optical Wavelength Conversion Schemes for Increased Input Power Dynamic Range", *IEEE Photonics Technology Letters*, vol. 10(1), January 1998.

[48] B. Lavigne, E. Balmefrezol, P. Brindel, B. Dagens, R. Brenot, L. Pierre, J-L. Moncelet, D. de la Grandilère, J-C. Remy, J-C. Bouley, B. Thedrez and O. Leclerc, "Low input power All-Optical 3R Regenerator based on SOA devices for 42.66 Gbit/s ULH WDM RZ transmission with 23 dB span loss and all-EDFA amplification", *Technical Digest, Optical Fiber Communication Conference 2003 (OFC 2003),* Atlanta, USA.

[49] T. Ito, Y. Shibata, A. Ohki, R. Sato and Y. Akatsu, "40 Gb/s Burstmode Optcal 2R Regeneration with Packet-to-packet Power Level Equalization". *In proc. of European Conference on Optical Communication 2004 (ECOC)*, Stockholm, Sweden, 2004.

[50] D. J. Blumenthal, B.-E. Olsson, G. Rossi, T. E. Dimmick, L. Rau, M. Mašanovic, O. Lavrova, R. Doshi, O. Jerphagnon, J. E. Bowers, V. Kaman, L. A. Coldren and J. Barton, "All-Optical Label Swapping Networks and Technologies", *IEEE Journal of Lightwave Technology*, vol. 18(12), December 2000.

[51] T. Fjelde, A. Kloch, D. Wolfson, C. Janz, A. Coquelin, I. Guillemot, F. Gaborit, F. Poingt, B. Dagens and M. Renaud, "Novel Scheme for Efficient All-Optical Label Swapping in Packet Switches using a Compact and Simple XOR gate", *Proceedings of European Conference on Optical Communication (ECOC 2000),* Munich, 2000.

[52] K. G. Vlachos, I. T. Monroy, A. M. J. Koonen, C. Peucheret and P. Jeppesen, "STOLAS: Switching Technologies for Optically Labeled Signals", *IEEE Communications Magazine*, vol. 41(11), p. 9-15, November 2003.

[53] C. W. Chow and H. K. Tsang, "Optical Packet Labeleling using Polarization Shift Keying (PolSK) Label and Amplitude Shift Keying (ASK) Payload", *In proc. of Optical Fiber Communication Conference (OFC'2005),* Anaheim, USA, 2005.

[54] T. H. Cormen, C. E. Leiserson and R. L. Rivest: *"Introduction to Algorithms"*, MIT Press, 1990.

[55] Fred J. Taylor, "Residue Arithmetic: A Tutorial with Examples", *IEEE Computer*, vol. 17(5), 1984.

[56] M. Dodge. "Maps of Internet Service Providers (ISP) and Internet Backbone Networks" *http://www.cybergeography.org/atlas/isp_maps.html* (2001) [online]

[57] M. W. Chbat, E. Grard, L. Berthelon, A. Jourdan, P. A. Perrier, A. Leclert, B. Landousies, A. Ramdane, N. Parnis, E. V. Jones, E. Limal, H. N. Poulsen, R. J. S. Pedersen, N. Flaarønning, D. Vercauteren, M. Puleo, E. Ciaramella, G. Marone, R. Hess, H. Melchior, W. V. Parys, P. M. Demester, P. J. Gødsvang, T. Olsen and D. R. Hjelme, "Towards Wide-Scale All-Optical Transparent Networking: The ACTS Optical Pan-European Network (OPEN) Project", *IEEE Journal on Selected Areas in Communications*, vol. 16(7), pp. 1226-1244, Sept. 1998.

[58] M. Faloutsos, P. Faloutsos and C. Faloutsos, "On power-law relationships of the Internet topology", *Computer Communication Review*, vol. 29(4), pp. 251-262, 1999

[59] A.-L. Barabsi, R. Albert and H. Jeong ,"Scale-free characteristics of random networks: The topology of the world wide web", *Physica A,* vol. 281(1-4), pp. 69-77, 2000.

[60] Y. Zhang, L. K. Chen and C. K. Chan, "A Multi-Domain Two-Layer Labelling Scheme for Optical Packet Switched Networks with Label-Swapping-Free Forwarding", *In proceedings of European Conference on Optical Communication 2003 (ECOC)*, Session Mo3.4.5, Rimini, Italy, 2003

[61] F. J. Taylor, "Residue Arithmetic: A Tutorial with Examples", *IEEE Computer,* vol. 17(5), pp. 50-62, May 1984

[62] R. Sivakumar and N. J. Dimopoulos, "VLSI architectures for computing X mod m", *IEE Proc.-Circuits Devices Syst.* vol 142(5), October 1995.

[63] S. J. Piestrak, "Design of Residue Generator and Multioperand Modular Adders Using Carry-Save Adders", *IEEE Transaction on Computers*, vol. 423(1), January 1994

[64] N. S. Szabo and R. I. Tanaka, "Residue Arithmetic and its Applicatons to Computer Technology", McGraw-Hill, 1967

[65] W. M. Wong and K. J. Blow, "Design of All-optical Processor for Novel Packet Forwarding Scheme in Optical Networks", *Proc. for Conference on Optical Internet 2002 (COIN)*, Cheju, Korea, 2002

[66] P. E. Green, "Fiber to the Home: The Next Big Broadband Thing", IEEE Communications Magazine, vol 42(9), September 2004.

[67] N. J. Frigo, P. P. Iannone and K. C. Reichman, "A View of Fiber to the Home Economics", IEEE Optical Communications, vol. 2(3), p. 16-23, August 2004.

[68] M. Weingarten and B. Stuck, "Is Fiber To The Home Affordable?", Business Communication Review, vol. 34(6), p. 24-28, June 2004.

[69] M. Abrams, P. C. Becker, Y. Fujimoto, V. O'Byrne and D. Piehler, "FTTP Deployments in the United States and Japan – Equipment Choices and Service Provider Imperatives", IEEE Journal of Lightwave Technology, vol. 23(1), p. 236-246, January 2005.

[70] ITU-T, G.992.5: "Asymmetric Digital Subscriber Line (ADSL) transceivers – Extended bandwidth ADSL2 (ADSL2+)", January 2005.

[71] ITU-T, G.993.1: "Very high speed digital subscriber line transceivers", April 2004.

[72] ITU-T, G.993.2: "Very high speed digital subscriber line transceivers 2 (VDSL2)", February 2006.

[73] J. Xie, S. Jiang and Y. Jiang, "A Dynamic Bandwidth Allocation Scheme for Differentiated Services in EPONs". *IEEE Communications Magazine*, vol. 42(8), pp. 32-39, August 2004.

[74] ITU-T, G.983.1: "Broadband optical access systems based on Passive Optical Networks (PON)", January, 2005.

[75] IEEE Standard 803.2ah: "*IEEE Standard for Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks--Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*", 2005

[76] Datasheet by NEL: "NEL Laser Diodes NLK5C5E3AA", July 2002. Available upon request.

[77] ITU-T, G.652: "Characteristics of a single-mode optical fibre cable", October 2000.

[78] ITU-T, G.694.2: "Spectral Grids for WDM Applications: CWDM Wavelength Grid", June 2002.

[79] O. Gauthier-Lafaye et al: "High temperature 10 Gbit/s directly modulated 1.3 μm DFB lasers using InAsP/InGaAsP materials", IEEE Electronic Letters, vol. 38, No. 6, March 2002.

[80] Datasheet by Go4Fiber.com, "Single Mode Standard Star & Tree: SSTS-1xN-NxN", http://www.go4fiber.com

[81] RBN, "CWDM, Technology, Standards, Economics & Applications", White Paper. [Online] Available: http://www.rbni.com/rbn_cwdm_tech_paper-1_20sep02.pdf

[82] Datasheet by Tsunami Optics: "4-Channel CWDM MUX or DMUX", 2002.

[83] Datasheet for optical circulator http://www.accelink.com/accelink/document/datasheets/29-cir.pdf

[84] Datasheet by GIGA: "10 Gbit/s Forward Error Correction, IXF30001", Revision S05, Nov. 2000.

[85] DS/EN 60825-2, European Standard: "Safety of laser products – Part 2: Safety of optical fibre communication systems.

[86] Datasheet by Bookham: "LC 131-98 10 Gb/s Directly Modulated Uncooled DFB Laser", July 2002.

[87] A. Veres, Zs. Kenesi, S. Molnár and G. Vattay, "TCP's role in the propagation of self-similarity in the Internet", *Computer Communications,* vol. 26(8), pp. 899-913, 2003.

[88] MUPBED Deliverable D0.1: "Project Summary", [online] http://www.ist-mupbed.org, September 2004.

[89] GÉANT2 website, http://www.geant2.net

[90] MUPBED Deliverable D2.2: "Preliminary Interface Specification", [online] http://www.ist-mupbed.org, February 2006.

[91] ITU-T: "International telephone connections and circuits – General Recommendations on the transmission quality for an entire international telephone connection", *ITU-T Recommendation G.114*, May 2003.

[92] GARA project website, http://www-fp.mcs.anl.gov/qos

[93] GRS project website, http://www.cs.ucl.ac.uk/staff/S.Bhatti/grs/index.html

[94] GARA based DataTAG project website, http://datatag.web.cern.ch/datatag

[95] Grid Just in Time Network project website, http://projects.anr.mcnc.org/Jumpstart/

[96] UCLP project website, http://www.canarie.ca/canet4/uclp/index.html

[97] B. Riddle, "BRUW: A Bandwidth Reservation System to Support End-user Work", *Proceeding of TERENA Networking Conference (TNC'2005)*, [online], http://www.terena.nl/events/tnc2005/programme/

[98] F. Baker, C. Iturralde, F. Le Faucheur and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 Reservations". *RFC 3175, Internet Engineering Task Force (IETF)*, September 2001

[99] J.T. Moy, "OSPF Version 2", IETF RFC 2328, April 1998

[100] D. Katz, K. Kompella and D. Yeung, "Traffic Engineering (TE) Extension to OSPF Version 2", RFC 3630, September 2003

[101] J.T. Moy, "OSPF – the Anatomy of an Internet Routing Protocol", Addison-Wesley, 1998.

[102] J.T. Moy, "OSPF Complete Implementation", Addison-Wesley, 2001.

[103] E. Mannie, ed., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.

[104] A. Fuqaha, "Routing in All-Optical DWDM Networks with Sparse Wavelength Conversion Capabilities", Proc. of GLOBECOM, vol.5, pp.2569-2574, December 2003

[105] OPNET Modeler Product Documentation, Release 10.5.